Object-focused Interaction in Collaborative Virtual Environments

Jon Hindmarsh[1], Mike Fraser[2],

Christian Heath[1], Steve Benford[2] & Chris Greenhalgh[2]

[1]King's College London

[2]University of Nottingham

Authors and Affiliations

1. Jon Hindmarsh (Contact Author)
Work, Interaction & Technology Research Group, The Management Centre, King's College London, London W8 7AH, UK.
Tel. +44 171 333 4194
Fax. +44 171 333 4479
E-mail: jon.hindmarsh@kcl.ac.uk

2. Mike Fraser
Communications Research Group, Department of Computer Science, University of Nottingham, Nottingham NG7 2RD, UK.
Tel. +44 115 951 4225
Fax. +44 115 951 4254
E-mail: mcf@cs.nott.ac.uk

3. Christian Heath
Work, Interaction & Technology Research Group, The Management Centre, King's College London, London W8 7AH, UK.
Tel. +44 171 333 4496
Fax. +44 171 333 4479
E-mail: christian.heath@kcl.ac.uk

4. Steve Benford
Communications Research Group, Department of Computer Science, University of Nottingham, Nottingham NG7 2RD, UK.
Tel. +44 115 951 4225
Fax. +44 115 951 4254
E-mail: sdb@cs.nott.ac.uk

5. Chris Greenhalgh
Communications Research Group, Department of Computer Science, University of Nottingham, Nottingham NG7 2RD, UK.
Tel. +44 115 951 4225
Fax. +44 115 951 4254
E-mail: cmg@cs.nott.ac.uk

ABSTRACT

This paper explores and evaluates the support for object-focused collaboration provided by a desktop Collaborative Virtual Environment. An experimental ‘design’ task was conducted and video recordings of the participants’ activities facilitated an observational analysis of interaction in, and through, the virtual world. Observations include: problems due to ‘fragmented’ views of embodiments in relation to shared objects; participants compensating with spoken accounts of their actions; and difficulties in understanding others’ perspectives. Implications and proposals for the design of CVEs drawn from these observations: the use of semi-distorted views to support peripheral awareness; representations of actions and pseudo-humanoid embodiments; and navigation techniques that are sensitive to the actions of others. The paper also presents some examples of the ways in which these proposals might be realised.

Keywords

Social Interaction, Virtual Reality, CSCW, Objects, Shared Spaces, Interfaces, Embodiment.

INTRODUCTION

Recent years have witnessed extraordinary advances in the quality and effectiveness of visualisation and VR technologies. Indeed, a wide variety of organisations and institutions are increasingly finding novel and innovative uses for these technologies, uses that encompass and include the fields of design, entertainment, medicine, engineering and so forth. However, most industrial applications are being used to support high-quality graphical visualisations of real (and imagined) scenes and settings, where individual users navigate around the world(s), whether on their own or with a local audience. However, in future years, these virtual settings and scenes could well become everyday work or meeting places for remote participants – for example architects discussing possible alterations to a design; or medical experts discussing and planning surgical techniques. Indeed, it is increasingly recognised that changes in the structure of contemporary organisations will place corresponding demands on technology to provide support for such distributed collaborative work [2]. The trend towards disaggregation, globalisation and dynamic networks of firms suggests that CSCW technologies will play an increasingly important part in supporting co-operation and communication amongst distributed personnel.

To provide adequate support for teamwork 'in' (and through) virtual environments, however, basic research is necessary to understand the kinds of resources and support individuals require to undertake seamless collaboration. In particular, it is widely recognised that much collaborative work rests upon the sharing and discussion of a whole host of documents, tools and other artefacts. This would seem of critical importance for the development of virtual environments where remote colleagues would need to discuss virtual objects and scenes and also collaboratively co-ordinate their navigation around, and looking within, the world.

Indeed, this would also seem to be a more general problem for the development of advanced shared workspaces. Although asynchronous text based systems to support remote work, such as e-mail, Notes and the World Wide Web, are flourishing within the business community, technologies to support real time, collaborative work, such as mediaspaces, have met with less success. These systems have not as yet proved to provide satisfactory domains for collaborative work, and even

their precursors, such as video-telephony and video-conferencing, have failed to have the impact that many envisaged. It has been argued that the relative weakness of many systems to support synchronous remote working derives from their inability to assist individuals in working flexibly with documents, models and other workplace artefacts [18].

In this paper we build on workplace studies and media space research to develop and evaluate a Collaborative Virtual Environment (CVE) designed to support real time collaboration and interaction around objects and artefacts. In particular, we wish to explore the extent to which the system provides participants with the ability to refer to, and discuss, features of the virtual environment. The implications of these observations are then drawn out with regard to the specific development of CVEs to support interaction around objects. We also consider more general issues relevant for the development of sophisticated support for distributed collaborative work.

BACKGROUND

In their wide-ranging investigation of organisational conduct, workplace studies have powerfully demonstrated how communication and collaboration are dependent upon the ability of personnel to invoke and refer to features of their immediate environment [e.g. 12, 15, 17]. Studies of settings such as offices and control rooms have shown that individuals not only use objects and artefacts, such as screens, documents, plans, diagrams and models, to accomplish their various activities, but also to co-ordinate those activities, in real time, with the conduct of others. Indeed, it is found that many activities within co-located working environments rely upon the participants talking with each other, and monitoring each others' conduct, whilst looking, both alone or together, at some workplace artefact. An essential part of this process, is the individual's ability to refer to particular objects, and have another see in a particular way, what they themselves are looking at [18]. These studies provide insights into the demands that will be placed on technologies that aim to provide flexible and robust support for remote working.

Interestingly, systems to support distributed collaboration are increasingly attempting to meet these needs. Rather than merely presenting face-to-face views, conventional video-conferencing systems are now often provided with a 'document camera', and media spaces and similar technologies are

increasingly designed to provide participants with access to common digital displays or enhanced access to the others' domain [e.g. 10, 16, 19, 26]. However, it is not clear that such systems provide adequate support for object-focused collaboration.

Our own earlier attempts to develop a media space to support variable access between participants and their respective domains or objects and artefacts have also met with limited success [10, 16]. In MTV II, an experiment undertaken with Bill Gaver, Abi Sellen and Paul Luff, we provided remote participants with various views of each other and their respective domains on three separate monitors, including a 'face-to-face' view, an 'in-context' view (showing the individual in the setting), and a 'desktop' view (allowing access to documents and other objects). Participants were asked to undertake a simple task which necessitated reference to objects in, and features of, each others' respective environment. Despite providing participants with visual access to the relevant features of each others' domains, participants encountered difficulties in completing the task. In general, individuals could not determine what a co-participant was referring to, and, more specifically, where, and at what, s/he was looking or pointing. This problem derived from participants' difficulties in (re)connecting an image of the other with the image of the object to which they were referring. The fragmentation of images – the person from the object and relevant features of the environment – undermined the participants' ability to assemble the coherence of the scene [15]. This undermined even simple collaboration in and through the mutually available objects and artefacts (for similar findings, see [1]).

In the light of these findings we decided to consider the kinds of support for object-focused work provided by CVEs. Of course, CVEs enable participants to work with shared access to objects located in the virtual environment, whilst media spaces endeavour to provide participants with the opportunity to work on 'real, physical' objects. However, we believe that CVEs may provide a more satisfactory method of supporting certain forms of distributed collaborative work. Firstly, the use of VR technologies in a range of pursuits could well involve collaborative work over and around virtual objects, designs and scenes in their own right. Secondly, CVE developers are refining techniques for integrating information from the physical world into virtual environments, for example in the form of embedded video views that are displayed as dynamic texture maps attached

to virtual objects [20]. Should CVEs prove to provide effective support for object–focused collaboration with virtual objects, then such extensions might allow them to provide similar support for remote collaboration with physical objects in the future.

Although for media spaces, the problems associated with establishing what another can see or is looking at are well recognised, it is often argued that problems of recognising what views and scenes are available to the other are 'naturally' overcome in 3-D worlds [24]. These would seem reasonable claims, especially because:

- even though the actual users are located in distinct physical domains, the CVE allows participants to share views of a stable and common virtual world consisting of the same objects and artefacts;

- the use of embodiments located in the virtual world, provides the participants with access both to the other, and to the other's actions and orientations in the 'local environment'. The embodiments can look at and refer to things and, thus, can be seen alongside the objects at which they are looking and pointing. In this way, and unlike media spaces, (representations of) participants are visibly 'embodied in', and 'connected to', the common world of objects.

The aim of our experiments is to assess the extent to which a CVE might actually support object-focused collaboration. Aside from this we are also interested in more general issues concerning how individuals interact and work with each other in virtual environments. Surprisingly, despite the substantial literature on communication in media space and parallel technologies, there is little analysis, either in CSCW or elsewhere, concerning the organisation of human conduct in collaborative virtual environments. Bowers, Pycock and O'Brien have begun to explore such issues. In [5], the interactional use of even simple embodiments is elaborated through a conversation-analytic approach to the study of activity through an early CVE system. Additionally, these authors explicate the ways in which actions within a CVE can relate to users' real-world conduct [6]. More recently, Steed et al. have conducted experiments in small-group behaviour [25]. This work utilises statistical analysis to suggest positive relationships between co-presence and presence, presence and immersion, presence and group accord, and also argues that some evidence exists that immersive-

style interfaces may confer leadership status to a participant within the remit of a task. The work presented in this paper, however, considers and concentrates upon how the visible properties of the virtual environment (i.e. objects others than the embodiments) feature in, and are often critical to the intelligibility of, the participants' actions and activities.

EXPERIMENTING WITH 'FURNITURE WORLD'

To investigate object-focused interaction in CVEs, we adapted a task from the previous studies of the MTV system [10]. Participants were asked to collaboratively arrange the lay-out of furniture in a virtual room and agree upon a single design. They were given conflicting priorities in order to encourage debate and discussion. The virtual room consisted of four walls, two plug sockets, a door, two windows and a fireplace (Figure 1) and we implemented this 'furniture world' using MASSIVE–2, a general purpose CVE platform that has been developed at The University of Nottingham [4].

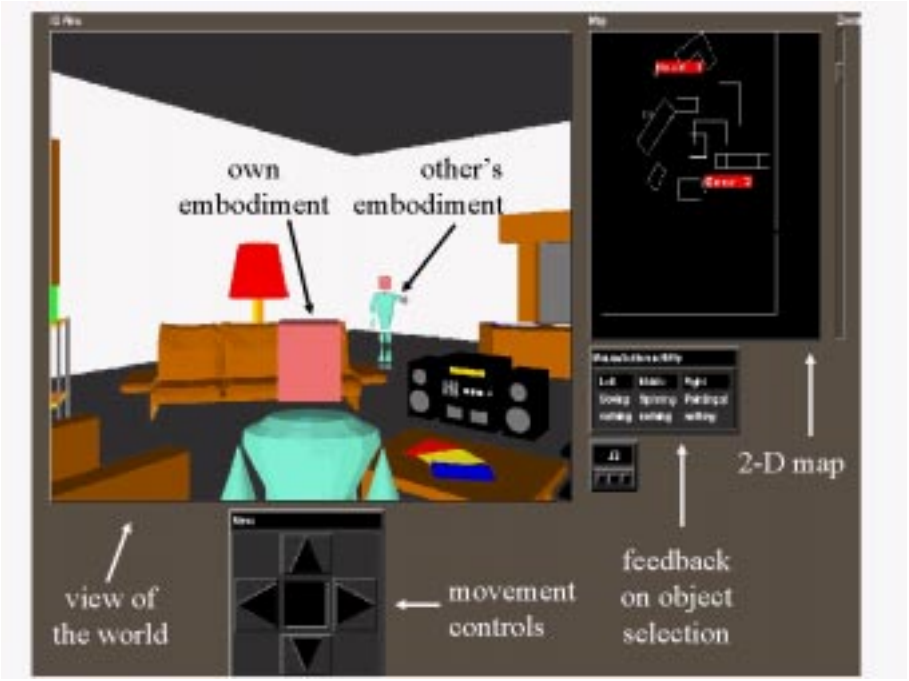
Figure 1: An Overview of Furniture World

MASSIVE–2 allows multiple participants in a shared virtual world to communicate using a combination of 3D graphics and real–time audio and to grasp and move virtual objects. The participants in our experiment used Silicon Graphics workstations connected via an Ethernet, with speech being supported by headphones and a microphone. It should be noted that the adaptation of a task from the previous media space experiments was not intended to assess which system provided the most adequate support. Indeed, a direct comparison would be rather deceptive as the technological differences make the task quite different – in MTV the participants have asymmetrical access to the model, whereas in the CVE all participants have equal access to the virtual furniture. Rather, this simple design task provides an opportunity to encourage (or even demand) that the participants discuss and discriminate features of the virtual world. Nevertheless, we have found it useful to reflect upon the differences regarding problems faced by the users of the two systems.

The Design of the Embodiment and Interface

In order to support our experiment, we extended the capabilities of the MASSIVE–2 default embodiment and simplified its default interface. The revised embodiment and interface are shown in Figure 2.

Figure 2: The Furniture World Interface

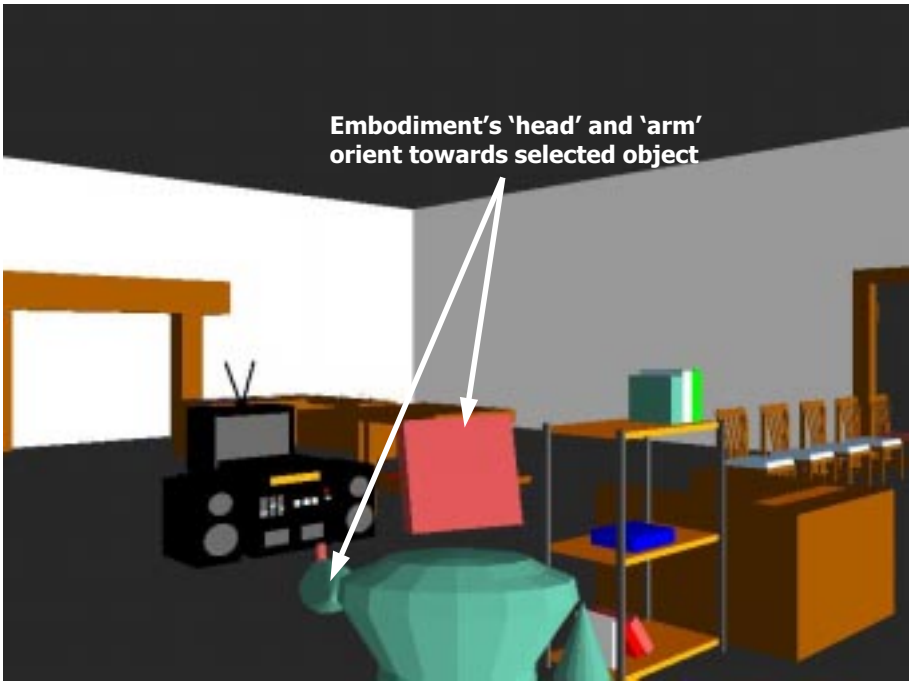


Our aim in revising the default embodiment was to enhance support for referencing visible objects. The guiding principle behind our design was that the embodiment should broadly reflect the

appearance of a physical body. Although photo-realism was not possible for performance reasons, this meant that the embodiment should be generally humanoid (i.e. have a recognisable head, torso and limbs). We adopted this approach because we felt that it is the most obvious choice and indeed, is one that has been widely adopted by CVE designers. One goal of our work was therefore to provide some insights as to the utility of pseudo-humanoid embodiments in CVEs.

Our embodiment supported pointing as a way of referencing objects. This was in addition to referencing them in speech or by facing them as were already supported by MASSIVE-2. A participant could choose to point at a target (an object or a general area of space) by selecting it with the right mouse button. Their embodiment would then raise its nearest arm to point at the target and would also incline its head towards it, as shown in Figure 3.

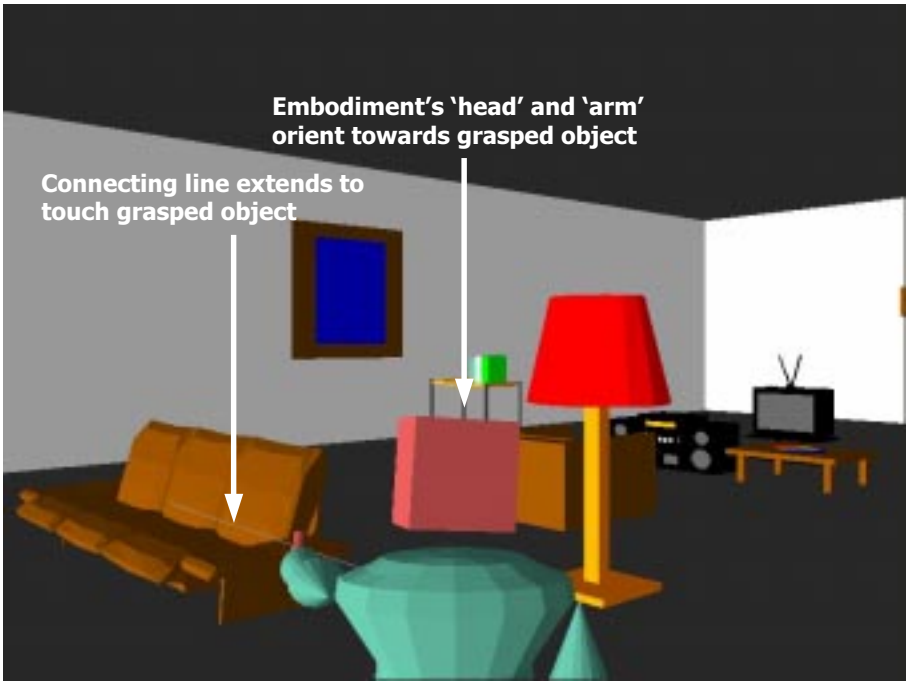
Figure 3: A User Points at the Stereo



Participants could grasp and move objects by selecting them with the left and middle mouse buttons. The left button moved the object across the floor of the room and the central button rotated it around its vertical axis. In order to simplify this particular design task, we removed the ability to lift objects off the floor and rotate them around other axes. This manipulation was also shown on the embodiment by raising an arm to point at the object and tilting the head towards it. In addition, a connecting line was drawn between the embodiment and the object being moved. This connecting

line was our only extension beyond normal physical embodiment. It was included to reflect the ability to manipulate objects at a distance in the CVE (e.g. without being within arm's length of them), as this is not a familiar experience in the physical world. An example of the user's view of the world whilst grasping an object is displayed in Figure 4.

Figure 4: A User Grasps the Sofa



In addition to this embodiment design, we took several steps to simplify the user interface. Participants could only carry out a limited set of actions. These were: *looking* (i.e., moving about on the ground plane so as to adjust their viewpoint), *speaking*, *pointing*, and *grasping* to move objects. Other simplifications included:

- restricting movement to the ground plane only (i.e. no 'flying').
- using of an out of body camera view that showed one's own body in the foreground of the scene. This technique was initially introduced in the MASSIVE system to extend one's field of view and to provide feedback as to the state of one's embodiment [13]. When pointing at an object with a viewpoint situated through the embodiment's 'eyes', for example, it is not possible to see the 'arm' move. Thus the only feedback that pointing is in progress is through the reporting facility on the interface. An 'out of body' view allows participants to see their own embodiment pointing and grasping. The field of view provided in our application was 55

degrees horizontally and 45 degrees vertically, in order to minimise distortion on the desktop interface.

Data Collection

Six trials of two participants and two trials of three participants were performed. Most participants were students, twelve male and six female, with a broad mixture of previous acquaintance. None of them had a background in CVE technology. Each trial took about an hour and consisted of ten minutes for participants to get used to the system, approximately half an hour to perform the given task, and then to conclude, we interviewed them about their experience of using the system.

A VCR was used to record each participant's on-screen activities and audio from their perspective – their own voice in real time, plus the other participant's voice(s) with a delivery lag over the network. Depending upon network traffic, the lag varied from being almost negligible up to imposing a one second delay on sound and image. Video cameras simultaneously recorded the participants in the real world (see the photograph in Figure 5) and contained audio from the participant in shot.

Figure 5: A User Being Filmed

The analysis of these recordings draws on conversation analysis [22] and focuses on a series of illustrative sequences extracted from, and resonant with other instances in, the data corpus. Conversation analysis has increasingly informed a range of studies (within CSCW, HCI and elsewhere) that have focused on the organisation of interaction in everyday workplaces [12, 17] and indeed in multi-user VR [5, 6]. We aim to contribute to this tradition.

It should also be noted that our use of a quasi-experimental approach is necessary for several reasons. Firstly, conversation analysis usually focuses upon naturally-occurring activities, but the incipient nature of the technology means that there are very few environments in which it is used as a matter of routine, except maybe by CVE designers themselves. Secondly, we know very little indeed about the organisation of interaction 'through' this communication medium and therefore it would be extremely premature to build hypotheses or to undertake large scale experimental studies. As a result, our approach is designed to explore and uncover the kinds of interactional phenomena, practices and problems that may be of particular relevance to both users and designers. In this way we aim to sensitise ourselves to the key issues that impact upon and engender the ways in which individuals interact and discuss objects in CVEs.

OBSERVATIONS

The participants found it relatively straightforward to accomplish the task asked of them. Indeed, they comfortably accomplished the desired task and even claimed to enjoy using the system. However, there are three key observations that would seem to have some import both for our understanding interaction in the CVE and for the development of distributed, shared workspaces:

- The image of an object under discussion is often 'fragmented' or separated from the image of the others' embodiment. This is primarily due to the narrow field of vision provided by the CVE (55°);
- Participants compensate for the 'fragmenting' of the workspace by using talk to make explicit actions and visual conduct that are recurrently implicit in co-present work;
- Participants face problems assessing and monitoring the perspectives, orientations and activities of the other(s) even when the other's embodiment is in view.

Fragmenting The Workspace

In this CVE (representations of) participants are located in a single, virtual domain. They are also given the ability to produce a very simple pointing gesture towards an object. This enables participants to use their virtual embodiments to indicate features within the common workspace. The following instance provides a simple example of how such gestures are used successfully to encourage another to look at an object with them. As we join the action, Sarah and Karen are repairing a confusion over which table should be moved (the initial in the margin indicates the current speaker – i.e. K for Karen and S for Sarah) .

Example 1: C20/2/98-14:29:45-VP:S

K: It's this table I'm talking about. this one yeah? *((K Points))*

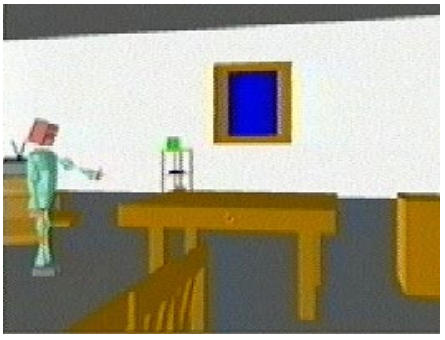
S: Yeah.

K: Can you see me point?

S: Yeah, it's the one in front of you isn't it.

Figure 6 shows the view that Sarah has of Karen's embodiment, her gesture and the table:

Figure 6: Sarah's View[1] [2]



Sarah is able to see Karen's gesture in relation to her surroundings. Moreover, she is able to see the pointing arm in relation to the relevant table. So, the ability to gesture is used as a successful resource with which to indicate an object in the world. The embodiment is located in the same domain as the referent, which enables the other to relate or 'connect' the gesture to the object referred to in the talk.

[1] For this one experiment, we did not provide users with a view of their own embodiment.

[2] This picture, as with the other illustrative pictures in this section, is taken from actual video data, and thus is subject to the resolution constraints of this medium.

A key finding from the MTV experiments, and one that would seem to resonate with many media space and shared workspace technologies, is that object-focused discussions are rendered highly problematic due to the 'fragmentation' of different elements of the workspace. For example, with an utterance such as 'in that corner there', the recipient may be able to see the speaker's face, the speaker's gesture and the relevant object. However, these are rarely seen together in one scene. They are often presented in different parts of the recipient's world [15]. For instance, the object could be at their fingertips, whilst the speaker's face is presented on one video screen and the gesture on another screen. Participants find it problematic to re-assemble the relations between body and object. As a result, they find it difficult to retrieve the referent.

In this CVE, the gesture and referent potentially can be seen alongside one another. Therefore, this type of fragmentation of the workspace is overcome, as shown in Example 1 above. Interestingly, however, a new version of the 'fragmented workspace' emerges. The 55 degree field of view provided by the desktop CVE restricts participants' opportunities to see embodiments in relation to objects. When an utterance is produced, individuals are rarely in an immediate position to see the embodiment alongside the referenced object. It turns out that it is critical that they do see them in relation to one another. So, they firstly turn to find the other's embodiment and then look for the object.

In Example 2, Sarah asks Karen about the 'desk-thing' in the room. Before they can discuss where they might put the desk, they need some 25 seconds to achieve a common orientation towards it (square brackets indicate overlapping utterances; a dot in brackets indicates a short pause in talk).

Example 2: C20/2/98-14:31:10-VP:K

S: You know this desk-thing?

K: Yeah?

S: Can you see- what I'm pointing at now?

((K Turns to Find S))

K: Er I can't see you, but [I think-

S: [Its like a desk-thing.

K: Er-where've you gone? [heh heh heh

S: [Erm, where are you?

K: I've- th- I can see

S: Tur- (.) oh, oh yeah. you're near the lamp, yeah?

K: Yeah.

S: And then, yeah turn around right. (.) and then its like (.) I'm pointing at it now, but I don't know if you can see what [I'm pointing at?

K: [Right yeah I can see.

When Sarah asks if Karen can see what she is pointing at, Karen starts to look for Sarah's embodiment and her pointing gesture. She is actually facing the desk very early on in the sequence, but ends up turning 360º, via Sarah's gesture, to return to the very same desk.

In co-present interaction, when an individual asks a co-participant to look at an object at which they are pointing, that co-participant can usually see them in relation to their surroundings. They simply turn from the body of the other to find the referenced object [18]. In this CVE, participants often do not have the other's embodiment in view during an utterance. They might turn out to have initially had the referent in view, but without seeing the object in relation to the pointing gesture, they have little information with which to assess if they are looking at the appropriate object; in other words, they may see a 'desk-thing', but is it the relevant 'desk-thing'? In some cases, then, they cannot be certain that they are looking at and discussing the same object without seeing that object in relation to the other's embodiment.

Participants find the relevant object by following a particular sequence. First they turn to find the gesture and then they use this as a resource to find the referent. Even in short and straightforward instances, participants can be seen to turn away from an object to find the other's gesture, only to subsequently return to face that object. Participants may, however, need to engage in an extensive search for their co-participant's embodiment before being able to see the relevant object.

These problems often arise because the other's embodiment is not visible at the onset of a new activity. However, misunderstandings can also arise when the other's embodiment *is visible*, but is again separated from the objects on which they are acting. For example, in the following instance,

Andre is turning around whilst suggesting possible design changes. He happens to rotate past Trevor's embodiment just as Trevor's virtual arm lifts up.

Example 3: B30/1/98-12:03:50-VP:A

A: I think maybe the couch can go in front of the

((T's arm rises))

-er fireplace. what you pointing at?

((A begins to rotate his view back towards T))

T: Just moving the telly a bit.

A: Oh right.

Andre curtails his own utterance in order to attend to the demands of Trevor's gesture and its potential relevance for him at this moment. Trevor is at the edge of his viewpoint (see Figure 7), so he cannot see the objects toward which Trevor seems to be pointing, so he asks "what you pointing at?".

Figure 7: Andre's View

The act of pointing is represented by an arm lifting and the head tilting slightly. The act of moving an object is represented in the same way on the embodiment. The only difference is that when the object is being moved, a thin line is projected out from the embodiment to the object. When Andre sees the embodiment (and its gesture) in isolation from its surroundings, he sees Trevor pointing and recognises that it could be designed for him. Unfortunately, Trevor is just re-positioning the T.V., but the line from the virtual arm is not thick enough, or in enough contrast with the background, to be visible.

Seeing the embodiment in isolation from the specific object on which it is acting, leads Andre to misunderstand Trevor's actions. This misunderstanding disrupts the onset of the new activity, namely a discussion over where to place the couch.

In co-present environments, the local complex of objects and artefacts provides a resource for making sense of the actions of others. The production of action is situated in an environment and its intelligibility rests upon the availability of that environment [15]. However, participants in CVEs only encounter a fragment of the visible world. Separating an embodiment from the objects on which they are acting creates difficulties for participants. Indeed, their overall sense of action is impoverished. As they are rarely in a position to see both object and embodiment simultaneously, they have problems in relating the two. Critically, the sense of talk or action is based upon the mutual availability of that relationship.

Even when participants have a referenced object in their view of the world, they need to see it in relation to the other's embodiment. Without seeing them together, they cannot be sure if it is the relevant object or not. In co-present interaction these resources are often simultaneously available. However, in this CVE participants have to follow a sequence of action to assemble the scene. First they must look for the other's embodiment, then they turn to find the object in relation to it.

Making The Implicit Explicit

The previous section highlights a problem related to the narrow field of view provided by the CVE. However, participants are sensitive to the possibility that the other is not in a position to see their embodiment or the objects on which it is acting. Therefore they use their talk to make explicit certain actions and visual conduct.

For example, in co-present interaction an individual may simply say "what do you know about this?" alongside a gesture. Their co-participant can often turn quite easily to see what 'this' is and attend to the query. In such a way, the referential action and the projected activity can be conflated [18]; that is, the presentation of the object and the initiation of the activity (e.g. asking a question) can be one in the same.

In the CVE, participants tend to engage in a prefatory sequence in which the identity of the relevant object is secured before the main activity continues. Typical utterances include “The thing that I’d like this room to have is erm (.) you see the fireplace which is like (.) there?” or “See this sofa here?”. The activity only progresses when the other has found the referenced object.

So, participants are sensitive to, and have ways of solving, the problems of working in a ‘fragmented’ environment. However, these ‘solutions’ do damage to common patterns of working – an added sequence is inserted into the emergent activity.

A clear illustration is Example 2, in which a 25 second search for the desk takes place prior to a discussion about where it could be moved. This problem is compounded by the slow speed of movement in the CVE, preventing quick glances to find the other and the object. Interestingly, these referential sequences can last longer than the very activities that they foreshadow – for example, the length of time it takes to establish some common orientation towards an object or location can be much longer than the simple query that follows.

Unfortunately, the additional time involved in establishing mutual orientation is not the critical concern. This prefatory sequence actually disrupts the flow and organisation of collaborative activities. In co-present interaction, participants are able to engage in the activity at hand, whilst assuming that the other can see or quickly find the referent. Within the CVE, participants become explicitly engaged in, and distracted by, the problem of establishing mutual orientation. Indeed, it becomes a topic in and of itself.

In the CVE, participants cannot assume the availability of certain features of the world and so attend to making those features available to the other. Rather than debating where an object should be placed or whether to remove a piece of furniture altogether, participants are drawn into explicit discussions about the character of an object. It inhibits the flow of workplace activity and introduces obstacles into collaborative discussions.

Interestingly, in the case of three-party interaction, it takes just one participant to say that they can see the referent, for the speaker to proceed. Speakers drop their pointing gesture and move the activity on. Unfortunately, this can leave the third party struggling to find the object when the

resources to find it (e.g. a gesture) have been taken away. They become restricted from participating in the emerging activity, or else they must interrupt the others to ‘catch-up’.

As well as greater attention to reference, participants use their talk to make explicit the visual conduct of their embodiments. In co-present interaction, when an individual points something out, they are able to see the movement of the other. That movement reveals if another is looking for the object and therefore engaged in this activity rather than some separate concern. It can also be used to establish whether they are in a position to see the relevant object.

In this CVE, often individuals are not able see their co-participant as they point something out. To compensate, their co-participants tend to ‘talk through’ what they are doing and what they can see. Consider Example 4, in which Trevor points out the door to Andre.

Example 4: B30/1/98-12:02:20-VP:T

T: Th-the door’s behind me.

A: Oh right.

T: Over here, can you see that?

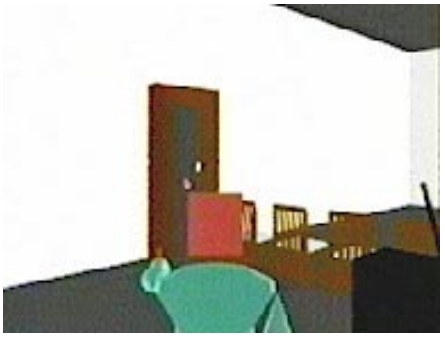
((T points towards the door))

A: I’m coming *((A rotates))*

T: Hang on *((T re-positions gesture))*

A: Yeah, okay, I got the door.

Figure 8: Trevor’s View



In pointing out the door, Trevor turns around and cannot see Andre (see Figure 8). Andre’s talk reveals certain critical aspects of his conduct to Trevor. For example, although Trevor cannot see

whether Andre is attempting to look for the door, Andre makes this available by saying “I’m coming”.

Given that movement in the world is relatively slow, participants often display that they are trying to look for the gesture and that the other’s actions are not being ignored. Often this is marked with phrases such as “Hang on, Hang on”, “I *am* looking” or even “errr” noises to fill the gap in talk. These actions would normally be available visually, through the sight of the other’s body movement.

When encouraging another to look at a particular feature of the local environment, participants attempt to design their referential actions for the other. In co-present interaction, they are even able to transform the course of a pointing gesture with regard to the emerging orientation of their co-participant(s) [18].

In this CVE, the problem for participants is that when they point to something, they often cannot see their co-participant(s). In co-present interaction, participants routinely configure a pointing gesture and then turn to view their co-participant’s ‘response’ [18]. This CVE does not allow participants to point at something and simultaneously look elsewhere. Therefore, it is much harder for them to be sensitive to the movements and visual conduct of the others’ embodiment. Their ability to design gestures or utterances to indicate an artefact is constrained.

This highlights a more general concern for participants engaged in collaborative work in CVEs. The organisation and co-ordination of much co-present work is facilitated by the ability to ‘monitor’ the activities of others. The narrow field of view, however, cuts out the visible features of many of those activities. So, participants’ ‘peripheral awareness’ of the other is severely constrained. The talk of participants does reveal features of their conduct. However, there is a much greater reliance on the talk than in everyday workplaces. Normally, individuals can rely upon the availability of the others’ visual conduct and see that visual conduct with regard to workplace artefacts. Here participants cannot. This leads to much cruder and less flexible practices for co-ordinating and organising collaborative work. Whereas they would normally be able to talk and simultaneously reveal other ‘information’ via visual conduct, almost all their actions must be revealed through talk.

Hidden Perspectives

Many of these examples have shown that participants face problems when the other's embodiment is not visible in their window on the world. However, even when the other's embodiment is visible, troubles often emerge. In particular, certain idiosyncrasies of the technology 'hide' how embodiments are viewing, and acting in, the world.

In the following instance, Pete is explaining to Rick where the fireplace is located. As he does, both Rick's embodiment and the fireplace are visible on his screen.

Example 5: D30/1/98-15:54:25-VP:P

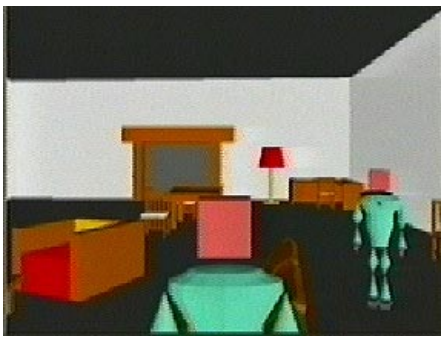
P: Do you reckon it might be better if we moved the T.V. over by the fireplace?
   (.)
R: By the fireplace?
P: Yeah [in the cor-
R: [Is there a fireplace in here?
   ((R Rotates))
P: In the cor- yeah you're facing it now.

Although Rick has the fireplace in his view, he does not recognise it as a fireplace. So when he says "is there a fireplace in here?", he simultaneously begins to rotate to his right to look for it. Pete treats Rick's embodiment as still facing or at least able to see the fireplace ("you're facing it now"). Unfortunately, at the moment he says this, Rick's viewpoint is focused on the door to the right of the fireplace[3]. Compare their viewpoints in Figure 9.

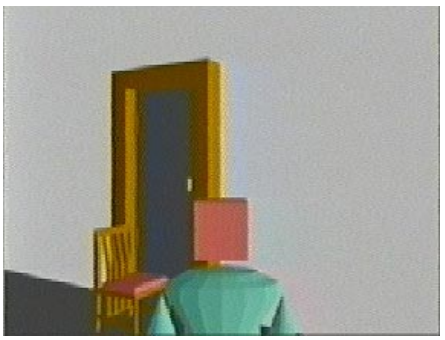
[3]In this case the delivery lag is negligible. In other instances, however, the lag disrupts the notion that this is a stable, common environment. When an individual says "now", for example, the other may hear it up to a second later. Therefore, if one is commenting on the other's actions, the other may hear those comments in relation to different actions than those for which they were produced.

Figure 9: A Comparison of Views

Pete's View



Rick's View



This reveals a problem for participants in assessing what the other can see. Even though they may have the other 'on-screen', it is hard for them to ascertain what is visible on that other's screen. In other cases, for example, participants assume the availability of their gestures, when the other is visible to them. Unfortunately, it turns out that the other cannot see them.

It may be that seeing a pseudo-humanoid form is confusing. This kind of embodiment may give participants a sense that it possesses a 'human-like' field of view, i.e. 180°. However, the users' field of view in this CVE is only 55°. Moreover, the CVE does not facilitate stereoscopic vision and the embodiments are often large virtual distances from their interlocutor(s), which exacerbates the problem, further concealing the other's perspective.

So, it is very difficult for participants to assess what the other might be able to see. This multiplies the problems raised in previous sections. It makes it far harder for participants to attempt to design actions for, and co-ordinate actions with, others. It is not simply that they need to get the other 'on-screen', because even then their sense of what the other is seeing is confused.

This issue also leads to problems with regard to the collaborative manipulation of objects – when one participant moves an object and the other directs that movement. When participants move an object in the CVE, their embodiment does not turn with it. Therefore, if they need to move an object out of the range of their field of view, they have to move it in stages – to the edge of their viewpoint, drop it, turn and then move it again. In Example 6, Mark and Ted are discussing where the standard lamp should be placed. Mark manipulates the lamp and Ted comments on his actions.

Example 6: A30/1/98-10:31:50-VP:M

M: Do you want the -er lamp moving?

T: Could do. see what it looks like.

M: Where d'ya want it put?

T: Erm- you decide

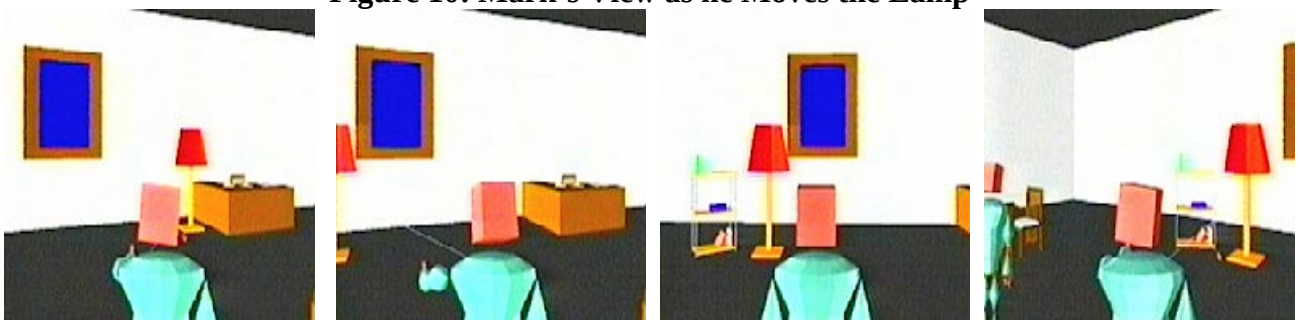
((pause as M moves the lamp))

T: That's right in front of the window, that's probably not so clever actually.

over the other side, over here.

Mark drops the lamp and Ted suggests that this is an inappropriate position for it ("that's right in front of the window, that's probably not so clever"). The video data suggests that Mark has not completed the movement and that this is a first 'step' in the positioning of the lamp. Notice in figure 10 that Mark places the lamp to the very edge of his screen before immediately beginning to turn to his left. However, the technology 'hides' this movement from Ted and at this point he rejects the positioning.

Figure 10: Mark's View as he Moves the Lamp



"T: That's right in front of the window, that's probably not so clever actually."

Ted cannot see how Mark is viewing and engaging with this activity. All he can see is the object being moved into an inappropriate position. So, he suggests an alternative.

Ted is prevented from getting a sense of how Mark views the lamp. He cannot see that the object is crossing the edge of Mark's window on the world. He is unaware that Mark cannot move the object

further without first temporarily releasing it and turning around. In such a way, the ‘boundary’ of Mark’s actions is hidden from Ted.

In everyday situations, if an individual finds an object too heavy or too awkward to carry, they can display the temporary nature of its setting-down through their visual conduct and demeanour. Bare virtual embodiments, on the other hand, conceal reasons for the placement of an object. Thus, the technology ‘disrupts’ access to the trajectory of moving objects. Objects that seem to have been placed once and for all, are often still in the process of being moved somewhere. This problematises an individual’s ability to assess what the other is doing, how they are orienting to the world and how they are engaging with the objects in that world. The technology thereby conceals critical aspects of the ongoing activity.

So, although an embodiment and the relevant object may be visible on screen, the relations of one to the other may not be visible or available. Moreover, the action is visibly produced in a different light to which it is seen and understood. This is important, because it makes it difficult for an individual to imagine ‘being in the other’s position’. As a result it is difficult to assess the nature or trajectory of the other’s actions. Thus, it is problematic for individuals to design and tailor their actions for co-participants, as they have little sense of how they are engaged in, or orienting to, the ongoing activity.

The orientations of the other in the world are hidden from view, which even leads to confusion about what they are doing. The technology distorts access to the common resources for working with others and making sense of their actions. Thus an individual’s understanding of the activity at hand can be disordered by the technology.

PRINCIPAL ISSUES

The CVE undermines certain resources commonly used to organise and co-ordinate collaborative work. Although the system does not prevent task completion (indeed, far from it), it does set up a range of obstacles to collaborative working. In particular, the system:

- reveals features of the world in ‘fragments’, due to its narrow field of view. This often separates views of the other’s embodiment from relevant objects. As a result, participants are provided with an impoverished sense of action. They cannot make sense of talk and activity without seeing (and seeking) the embodiment in relation to relevant features of the environment;

- forces participants to compensate for the problems of interacting in a ‘fragmented workspace’ by explicitly describing actions and phenomena that are unproblematically available in co-present interaction. In particular, referencing visual features of the world becomes a topic in and of itself, rather than being subsumed within the more general activity;

- reduces opportunities for participants to design and co-ordinate their actions for others. (Peripheral) awareness of the actions and orientations of the other is significantly undermined. Even when the other’s embodiment is visible on-screen, the technology disrupts the resources used to make sense of an individual’s activity.

For co-present interaction, Schutz suggested that we assume our different perspectives on the world and on an object are irrelevant for the activity at hand [23]. Individuals have a relatively sound understanding of what a colleague can see within the local workspace. Indeed, research suggests that individuals exploit their colleagues’ visible orientations in order to initiate new activities or to collaborate on particular tasks [15, 17]. The technical constraints imposed by the CVE render such activities more problematic. This is likely to lead to more intrusive means of monitoring others’ actions and interleaving activities with them. If CSCW technologies do not wish to impede the expedient production of work in the modern organisation, it is suggested that these issues are important for the design of systems to support synchronous remote working.

IMPLICATIONS

These observations lead us to conclude that certain technical issues should be addressed if CVEs are to provide more robust support for distributed collaboration. We propose that four key limitations have contributed to the phenomena noted above. These are:

- Limited horizontal field of view – it is difficult to simultaneously view the source (i.e. embodiment) and target of actions such as pointing and looking and confusion arises due to the difference between actual field of view for a participant and that anticipated by observers.
- Lack of information about others' actions – not all actions are explicitly represented on embodiments or target objects. Where they are, it may not be easy to distinguish between them.
- Clumsy movement – movement in CVEs may be slow due to problems of locating destinations, controlling the interface and system performance.
- Lack of parallelism for actions – the interface disallows some combinations of actions from being performed concurrently (e.g. moving and grasping; pointing and looking around).

For the remaining sections of this paper, we illustrate some of the ways in which CVEs might address these issues. The examples presented will include possible developments of the system, but the re-analysis of those developments for the task is not discussed here. Rather, the ways in which the CVE could be enhanced to provide support for both this and other kinds of object-focused co-operative work will be examined.

The above limitations suggest a range of general solutions. One approach is to replace the conventional desktop computer with a more immersive interface that would provide a wider field of view, more rapid movement and greater parallelism of action. Head–mounted displays (HMDs) might enable more rapid movement within the virtual world, especially glancing left and right. When used in conjunction with multiple position sensors, they might increase parallelism of action, for example, supporting simultaneous two–handed interaction with head movement. One could grasp one object and point at another while looking around. On the other hand, the field of view of all but the most expensive HMDs is very limited, although this may be compensated by the ability to rapidly glance around. Of course, HMDs introduce other problems: they are cumbersome, fragile and often low–resolution. They have yet to see widespread use or to emerge as a mass–market interaction device.

Projection–based systems provide another route to immersion. The most extreme example is a CAVE, a purpose built framework that completely surrounds a user or small group of users with

multiple synchronised back–projected views of a virtual world [7]. CAVEs fill the user's field of view and support unencumbered movement. Stereo projection and interaction using various 3D devices can further enhance the display of the virtual world. However, like HMDs, CAVEs introduce their own problems, especially their high cost and physical space requirements.

While recognising the potential of immersive displays such as HMDs and CAVEs to address the above limitations, the remainder of this section focuses on how we might improve the desktop CVE interface that was described earlier. This is because we expect desktop displays to remain the dominant form of CVE interface over the next few years. Even if we anticipate some improvements to the desktop interface such as the introduction of wide–screen displays, the above limitations will largely remain. We begin with the problem of field of view.

Increasing field of view with peripheral lenses

It has previously been mentioned that the desktop CVE interface provides participants with a horizontal field of view of approximately 55 degrees. This value, approximately a third of a human–like horizontal perceptual range, is typical of a desktop rendered viewpoint on a virtual environment. Although it is easy to widen the field of view in the rendering software, this results in extreme perspective distortions when displayed on a conventional narrow monitor. We anticipate that such distortions would make it difficult for participants to interact within the virtual world (although this remains to be proved). Indeed, distortions of the kind necessary, for example, to provide a human-like field of view would disrupt 'familiar' features of action in real world domains, which is a key reason why this kind of approach is generally avoided by CVE designers. For example, the ability to point or move directly toward something without veering away from it or continually adjusting your trajectory would be important, as indeed would the ability to assess the trajectory of someone else's actions. Moreover, particular activities that could be supported by virtual worlds will demand that participants have clear views of virtual objects. For example, what would be the point of collaborative design discussions in virtual worlds, if the virtual scene was distorted or transformed in order to support interaction? In the collaborative environment, the designers would be seeing distorted views of their designs-in-progress, thereby impeding the discussions about possible changes and so forth.

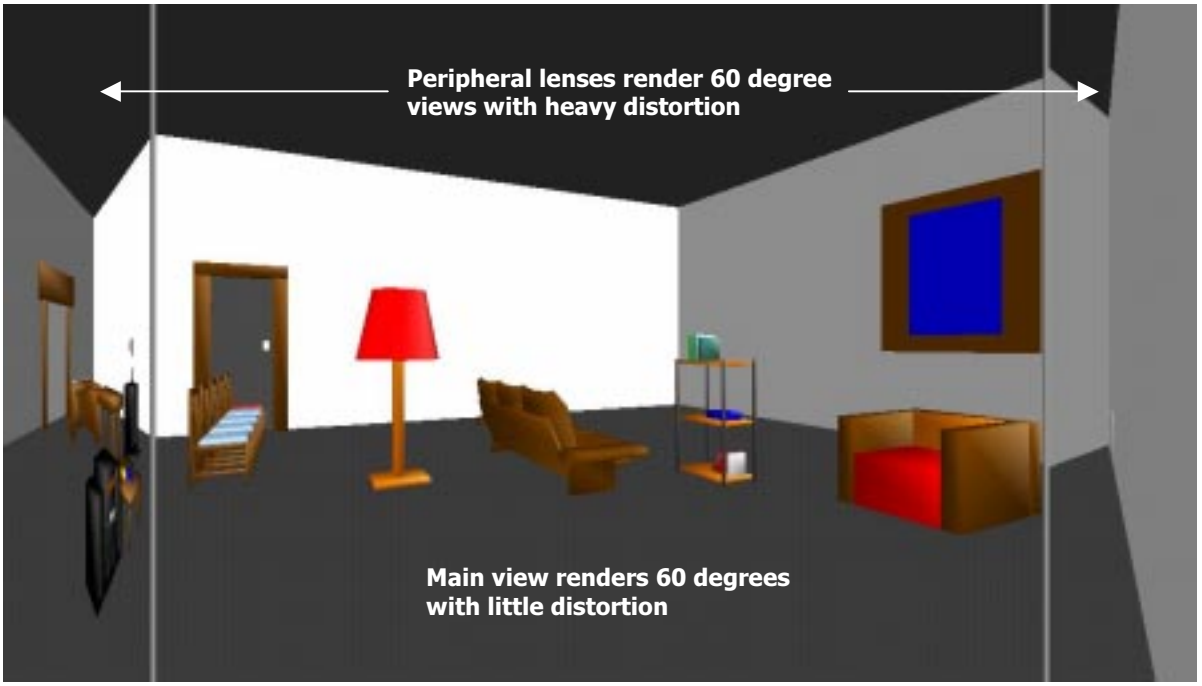
One technique we have previously employed within our own systems is to provide participants with different camera viewpoints in relation to their embodiments, in order to allow a 'framing' of the scene, or activity at hand [13]. For example, locating the camera behind their embodiment looking over their own shoulder allows participants to see themselves within the scene and gives a wider perspective than a strictly first person view. Some computer games have extended this with automated camera viewpoints in relation to the focus of activities within the 'task'. However, the success of automated viewpoints may be tied to the activity being supported – after all, racing in a driving simulation is quite a different pursuit from discussing objects with a colleague and making your actions intelligible. Indeed, a combination of providing participants with the ability to, and algorithmically causing the virtual interface to, frame activities from constantly varying perspectives might well prove problematic in maintaining a reciprocity of perspectives. Even providing some viewpoint feedback may not help – our related work on the MTV-1 system revealed that the feedback monitor was not intuitively used to assist object-focused discussions [10, 16].

Another possibility is to introduce more controlled perspective distortions than would be obtained by just widening the field of view in software. Techniques for utilising perspective distortions such as fish–eye lenses [8] could be applied for interfaces to virtual environments. A similar approach is the idea of *peripheral lenses* as introduced by Robertson et al. [21]. Their implementation consists of two additional windows that render views on a virtual environment to the left and right of the main view, but with increased distortion allowing more information to be rendered within a smaller horizontal space. The main view remains undistorted. This technique was primarily introduced as a navigation aid, thus their quantitative analysis focused on a single user search task in a virtual world.

Their conclusions stated that search times were not statistically improved by the use of peripheral lenses. We suspect, however, that an implementation based on this approach might prove more successful at supporting peripheral awareness in collaborative situations. In particular, studies of work in scenarios such as London Underground control rooms and financial dealing rooms suggest that "'Peripheral' monitoring or participation, appears to be an essential feature of both individual and collaborative work within these environments" [17]. Therefore, we have extended our CVE

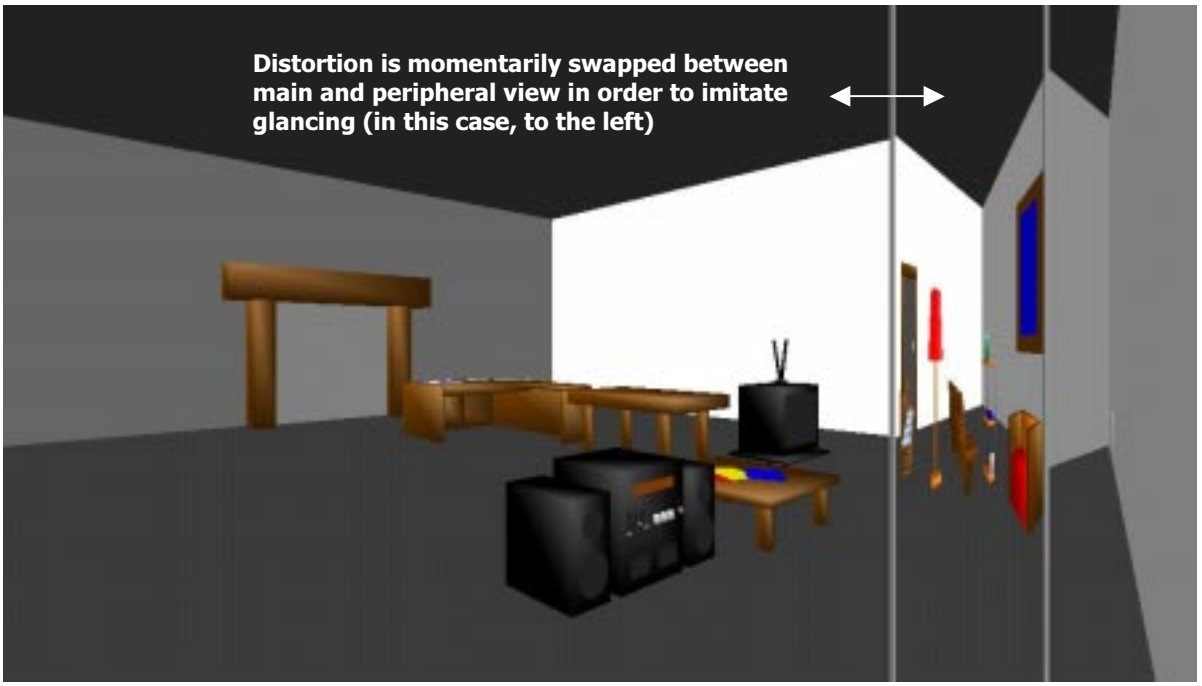
interface to make use of peripheral lenses. The desktop computer displays employed in the trials measured 12 inches horizontally. Our implementation allocates 2 inches for each peripheral lens, each of which renders a 60 degrees field of view utilising heavy perspective distortion. We allocate 8 inches for the main view and this also renders a 60 degree field of view, but with little distortion. Figure 11 displays this example.

Figure 11: View of Furniture World Using Peripheral Lenses



Our implementation extends peripheral lenses with the ability to focus on either periphery, through a technique we coin ‘peripheral glancing’. The user can momentarily swap their focus of attention to either the left or right peripheral view. Depressing a large button situated below the peripheral lens with a mouse enables a peripheral glance. The relevant lens is then widened so that it becomes undistorted and the main window is correspondingly narrowed. Releasing the button returns the main window and peripheral lens to their original condition. It is hoped that some of the problems, which might occur in misperception of other’s activities through heavy visual distortions, be avoided through the use of this glancing facility. An example of a user ‘glancing’ to the left is shown in Figure 12.

Figure 12: 'Peripheral Glancing' in Furniture World



Representation of actions and pseudo–realism in embodiment

Our second proposal focuses on the issue of providing better information about others' actions. Developers of 3-D spaces tend to represent actions on the source embodiment alone (e.g. raising an arm to show pointing), but, given the limited field of view available on desktop CVEs, our observations reveal that the source embodiment and target object are rarely simultaneously in view. Therefore, participants often find it difficult to find the object being discussed or referred to. Thus, we propose that the pseudo–realistic approach of showing actions solely by moving the source embodiment is too understated. Indeed, as well as widening an individual's view of the world (with peripheral lenses), we intend to support 'awareness' through visibly 'embedding' the actions of participants in the general environment. So, we have chosen to extend the representation of actions to include the source, target and the intervening environment. Our aim is that an action should be visible even if its observer has only one of these in view at the time.

Representing actions in the intervening environment may involve 'in–scene' representations, such as drawing a connecting line between the source and target, or 'out–of–band' representations such as playing back an audio sample or labelling the action on the observer's view in the style of a head–up display. Furthermore, representations should reveal not only which bodies and objects are

related by actions, but in what manner they are related. With this emphasis, an observer can get a sense of the other's actions from seeing, in isolation, their embodiment, the object that is being acted upon, or even the space between the two.

We propose that the representation of actions within CVEs should be designed according to the following three steps:

1. Identify all general actions – this includes actions such as 'looking' and 'speaking', as well as more explicit gestures such as pointing.

2. Identify the targets of each action – some actions such as grasping an object address an explicit target, whereas others such as speaking address a number of objects and embodiments populating a region of virtual space.

3. Determine whether and how each action is represented on the source embodiment, target object(s) and in the intervening environment – make sure that these representations are both consistent and distinguishable.

Retrospectively applying this approach to our initial interface design results in the Table I.

Table I: Action Representation Matrix

Action	Targets	How represented on embodiment?	How represented on targets?	How represented in environment?
Look	region of space	orientation of embodiment	no representation	no representation
Speak	region of space	no representation	no representation	voice on audio channel
Point	object or region	raise single embodiment's arm	no representation	no representation
Grasp	object	raise single embodiment's arm	target objects may be seen to move	connecting line between arm and target object

Four actions were possible: looking (moving one's embodiment), speaking, pointing and grasping. Our table shows that several actions were not represented on one or more of the source embodiment, target and environment and that other actions shared a common representation. For example, a single raised arm was used to show both pointing and grasping on the embodiment. Only one action, grasping, could be seen on the target, and only then when it was actually being moved. Also, there were no representations of looking or pointing in the environment. Table II is a revised

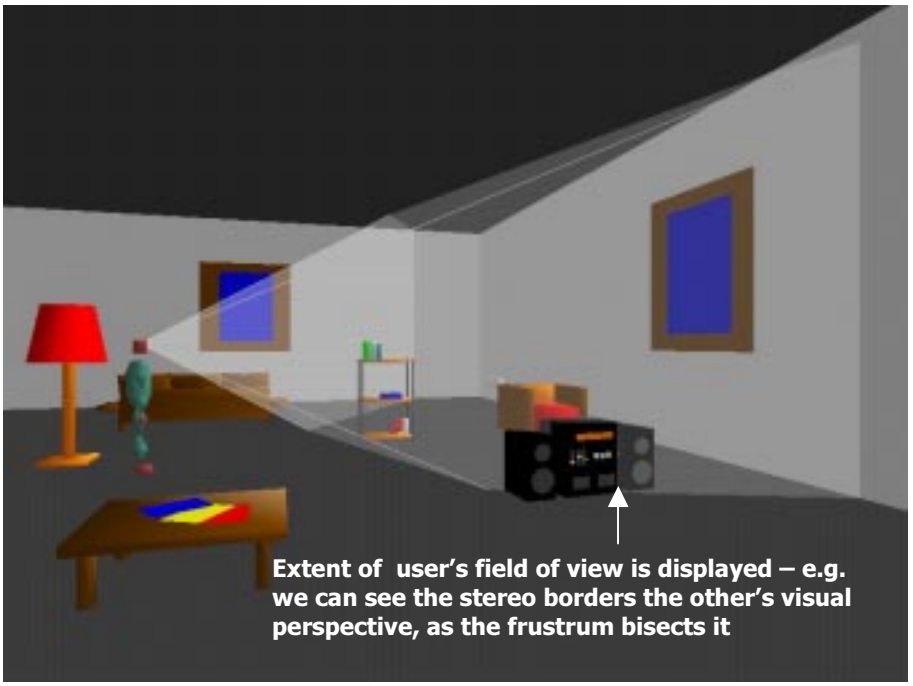
matrix indicating a number of extended representations of these actions (changes are shown in bold type).

Table II: Revised Action Representation Matrix

Action	Targets	How represented on embodiment?	How represented on targets?	How represented in environment?
Look	region of space	orientation of embodiment	**subtle highlight or shadow**	**visible display of view frustrum**
Speak	region of space	**speech bubble when audio sent**	**ears appear/enlarge when audio received**	voice on audio channel
Point	object or region	Raise single embodiment's arm	**highlighting the target object or region**	**visible ray of light between arm and object**
Grasp	object	**Raise both embodiment's arms**	**wireframe target seen to move**	**extend both embodiment arms to the object**

We can extend the represention of looking to include its targets, perhaps through subtle highlights or shadows on objects in view, and the environment, in this case by making the embodiments view (the extent of its field of view) visible as a semi-transparent frustrum. This second extension is shown in Figure 13 where the other participant's view is made visible. Future research will explore the benefits and problems associated with different representations of view frustra, how to incorporate additional representations for displaying peripheral lenses and glancing, and whether the use of lighting is sufficient or other techniques are more effective.

Figure 13: Visible View Frustrum in Furniture World

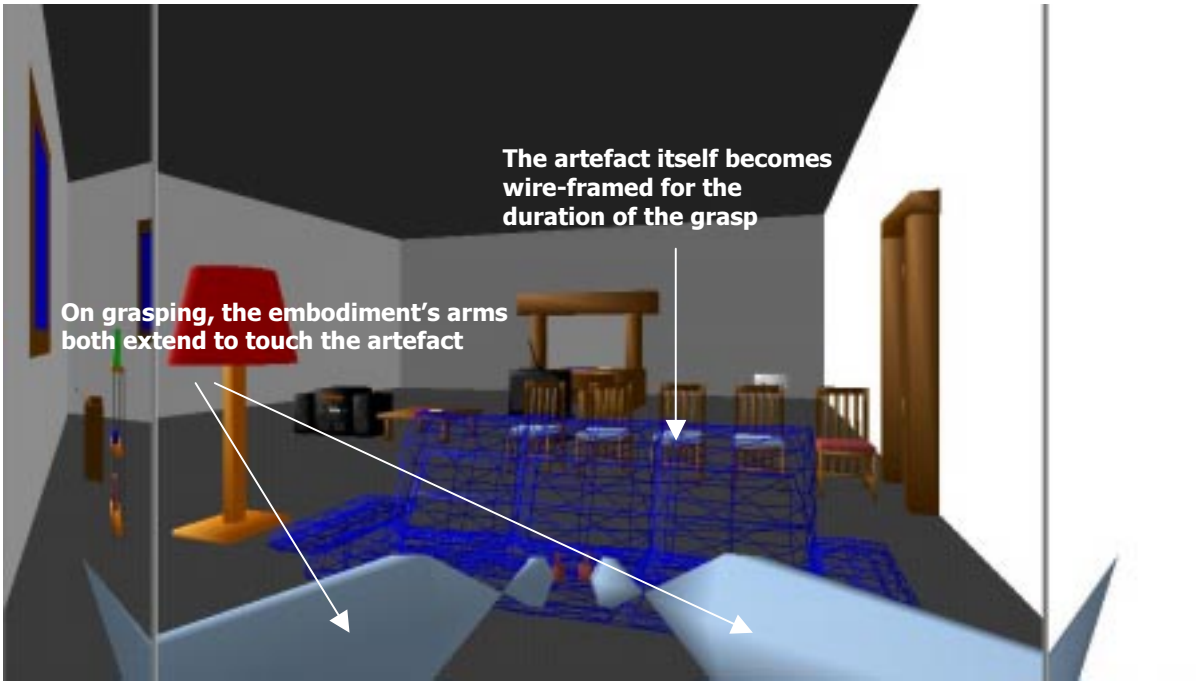


We can extend the representation of speaking to include both the embodiment and the target. The former might be represented as a speech bubble that appears above the embodiment's head whenever the user speaks. In fact, this is a standard feature of MASSIVE–2 that was switched off in the initial experiments as they mainly featured only two participants. With a greater number of individuals in a virtual world, assessing who is talking at any one time becomes a greater problem. MASSIVE–2 uses a silence suppression algorithm so that audio is only transmitted when the user speaks with sufficient volume in order to reduce network traffic and this also triggers the appearance of the speech bubble. Representation of speaking might be shown on target embodiments through ears that appear whenever an embodiment receives audio. A similar idea was included in an early embodiment design in the DIVE CVE platform; this embodiment's ears would appear whenever the embodiment was capable of receiving audio from the observer (i.e., was in virtual hearing range) [3].

We can extend the representation of pointing to include target objects and the surrounding environment. The former involves highlighting the targets. The latter involves rendering a visible ray of light between the embodiment and this target.

Finally, we suggest extending the representation of grasping in several ways. First, we distinguish grasping from pointing on the embodiment by raising two arms for the former and one for the latter. Second, we show the target as wireframe when it has been grasped, even if it is not currently moving. Third, we show grasping in the environment by extending the embodiment's arms to reach through the intervening space and touch the object (as if they were pieces of elastic attached to it), clearly differentiating them from the ray of light used for pointing. The following figure shows an example of representing grasping in this way (it is taken from the perspective of the grasping embodiment).

Figure 14: Grasping and Wire-framing an Object with Elongated Arms



It may be that other 'key' actions could be displayed in helpful ways. It is worth noting at this point that peripheral awareness need not be associated simply with visible resources. Audio signals could equally be used to make manifest certain actions. For example, currently participants often reveal their movements to look for an object through talk (e.g. "I'm coming", "hang on, hang on", etc.). Therefore, it would be worth investigating if suitable sounds could helpfully be used to display such conduct 'automatically'.

These issues relate to the general design rationale for embodiments in CVEs. Generally within CVE design, there is an aim to use 'realistic' or 'humanoid' embodiments [e.g. 14]. Indeed, some may argue that the embodiments used in this experiment were not realistic enough and that this was the source of users' problems. However the findings would suggest that the straight-forward translation of human physical embodiments into CVEs (however detailed) are likely to be unsuccessful unless participants can also be provided with the perceptual capabilities of physical human bodies including a very wide field of view and rapid gaze and body movement. Indeed, to compensate for the curious views on, and possible behaviours within, virtual environments, more exaggerated embodiments are required. Moreover, actions may best be represented across disparate features of the environment to enhance access to the sense of an action in the CVE.

Navigation based on others' actions

We propose the adoption of a form of target–based navigation, where the user is offered shortcuts for moving or turning towards relevant targets, as a way of dealing with clumsy navigation in CVEs. Furthermore, in a co–operative application, the choice of targets should depend to a large part on other participants' actions. For example, if an observer is looking at an embodiment that is pointing at an object, then the target of this point becomes a likely destination for the observer's next movement. Conversely, if they can see the target, then the source embodiment becomes a possible next destination. Such a mechanism would make it easier to glance back and forth between embodiments and the targets of their actions. Different actions might carry different weights for prioritising targets. For example, grasping an object or pointing at it might be seen as more significant than just looking at it. A more densely populated CVE might require more sophisticated techniques for selecting targets, for example taking account of the number of people who are acting on a target (e.g. looking or pointing at it) when determining its priority – a 'popular' object would become a high priority target.

Once suitable targets have been determined, they have to be presented to the user so as to enable their selection. They might be offered as lists of names or icons on their display. Alternatively, the ability to glance left or right, as introduced for peripheral lenses, might be extended to also focus on the nearest target in the specified direction. However, the specific ways in which alternative targets are represented and chosen need to ensure that participants are not overly drawn into concerns about what view to select, but rather support involvement in the activity and task at hand. Indeed, the danger in presenting various views may be that participants are alienated from their involvement in the task at hand by virtue of their need to switch between, and select, different views (for related discussion and examples, see [11]).

The system also has to manage the ways in which a switching between views is displayed to users. Here, we aim to learn from problems noted by Kuzuoka et al. [19], who discuss experiments with the GestureCam system in which instructors commented on tasks by remote participants. They had a camera sited in the remote setting, but often the narrow field of view of the camera made it hard for instructors to see certain objects. Of particular interest here, is that when instructors asked for

the camera to be moved, “the instructor often lost track of his position” [19: 40]. So, our proposal for the CVE is that movements between different views on the world should involve an animated transition, both from the point of view of the user and of co-participants. This would attempt to preserve the fluidity of the co-operative interaction and maintain the users’ overall sense and knowledge of the space.

More generally, we suggest that support for navigation, and co-participation, in CVEs could be oriented to the activities of others rather than with sole regard for the individual’s current actions. A point should be raised here about the nature of targets of activities. Looking, pointing and speaking may have targets for activity which are difficult for the CVE application to determine (for example, looking at a region, pointing to an area, talking to a group of embodiments). However, we might exploit models of user awareness such as the spatial model of interaction [3] that allow users to control the shape of their attention to and projection of information within a CVE using the mechanisms of focus and nimbus. These mechanisms are typically defined in terms of spatial regions rather than individual objects and allow the system to compute a level of awareness between each participant and each object (including other participants) in the environment.

Supporting parallel actions

Our final observation concerns the need to allow greater parallelism of action in the CVE. This might be possible to some extent when using a single mouse, for example, allowing the user to pick–up an object and put it down as separate actions that might be interleaved with other actions such as moving and pointing while they were ‘carrying it’. Greater improvements might be possible through the use of multiple input devices, for example using joysticks and other standard 3D interaction technologies alongside a mouse and keyboard.

Note on design for activities and applications

It must be stressed that these proposals are examples of how actions might be more explicitly represented to support the activities involved in this experimental design task. Although we hope to have raised issues that will have generic import, we expect that actual mechanisms will be highly activity dependent. In particular, several factors will have to be borne in mind when choosing representations and support for actions in CVEs.

Firstly, crowded environments that involve many participants may require additional mechanisms for limiting the use of such representations. For example, extended representations of actions may be used only for the most proximate or relevant participants or at the highest level of detail. Imagine highlighting every object that was in anybody's field of view in a crowded CVE.

Secondly, some actions occur much more frequently than others and some, such as looking, occur continuously. Such actions will require very subtle portrayal if the environment is not to become cluttered with additional information. To this end, designers could also consider which actions are key to the production of particular activities.

Thirdly, and relatedly, the design should be sensitive to the organisation of the activities involved in a particular application. At a gross level, different applications may demand different kinds of representations for key actions. For example, if the participants were attempting to discuss and discriminate different features of an object (e.g. surgeons discussing a virtual body), then wire-framing would be less useful for indicating parts of that object or ethereal features of common objects. Instead, more subtle means of indicating particular features and views of objects would be required. More likely, however, is that support for actions should be flexibly available to reflect the differing demands of the various activities relevant to any application. This is not, however, an easy issue. How such flexibility of design can be built in to the system, and how different kinds of resources and representations can be made available to participants at different times in their interaction is very much a matter for future research. Nevertheless, it is a critical issue in providing robust, but flexible, support for collaboration in VR.

CONCLUSION

We have explored how CVEs might support collaboration that is based upon the sharing of objects and artefacts. We carried out a study of object–focused collaboration with a typical CVE configuration, running on a standard desktop computer, using a mouse–based interface and representing the participants as pseudo–humanoid embodiments. We included the ability to point at and grasp and move objects. Our analysis of participants' communication within this environment raised three key issues. First, participants were able to make reference to objects in the shared

environment through pointing gestures. However, problems of fragmentation were observed. For example, there were difficulties resolving pointing gestures when the pointing embodiments and target objects were not both in view as was often the case. Second, participants compensated for this fragmenting of the workspace by using talk to make available certain actions and visual conduct, actions that are recurrently implicitly and unproblematically available in co–operative work. Third, participants faced problems assessing and monitoring the perspectives of others and establishing a sound sense of what they could see.

We have argued that a number of limitations in the technology have contributed to these problems. These include the narrow field of view offered by the CVE interface, the difference between this field of view and that anticipated by observers (who might assume that a humanoid embodiment has a human–like field of view), slow and difficult movement, and problems with carrying out actions in parallel.

In response, we have proposed a number of extensions to our CVE interface. Distorted peripheral lenses might increase the field of view. The use of multiple input devices might enable greater parallelism. We have also proposed that we could break with the approach of designing strictly humanoid embodiments and instead focus on exaggerating the representation of actions so that they can easily been seen by others. In particular, each possible action on the CVE might be represented on the source embodiment, the target object(s) and in the surrounding environment. Finally, we have proposed developing new navigation techniques that offer participants shortcuts such as glancing to nearby objects. We have proposed that the choice of shortcuts should be based upon other participants' actions within the environment, for example, it should be easy to turn towards an object that someone else is pointing at or grasping.

Our paper has focused on CVEs. However, it is interesting to consider how our proposals might be applied to other collaborative technologies, for example video–based technologies such as media spaces and video conferencing. Video cameras have a different (usually smaller) field of view than humans. Perhaps we should explicitly represent the view frustra of video cameras within physical environments so that people can more readily understand the extent of a remote viewers perception. We might mark them on the floor using carpets or imply them through suitably shaped furniture. As

another example, in a video system that uses tiled video windows showing the outputs of multiple cameras, we might offer the user hints as to which windows to look at or bring to the foreground based upon other users' actions. If I look at a particular target by bringing its video window into the foreground or increasing its resolution or size, for example the view from a document camera, then it might become a target for you to do the same.

We close with a final observation. There is long–standing discussion in fields associated with the analysis of collaboration in technology as to the ways in which social science, and in particular naturalistic studies of work, can inform the design and deployment of complex systems. Less attention is paid to the contribution of systems design to social science. The materials discussed here raise some potentially interesting issues for studies of work and interaction. In particular, the analysis of interaction in CVEs, like earlier discussions of media spaces and MTV, point to critical, yet largely unexplicated aspects of collaborative work. In particular, they reveal, *par excellence*, how collaborative activity relies upon the participants' mundane abilities to develop and sustain, mutually compatible, even reciprocal, perspectives. Critically they also uncover the resources on which participants rely in identifying and dealing with incongruities that arise. Whatever our sensitivities about using 'quasi–experimental' data, they provide, as Garfinkel suggests, 'aids to a sluggish imagination' [9]. They dramatically reveal pre–suppositions and resources which often remain unexplicated in more conventional studies of the workplace. Whilst we believe these pre–suppositions and resources are of some importance to sociology and cognate disciplines, it can also be envisaged how they may well influence the success or failure of technologies designed to enhance physically–distributed collaborative work.

Acknowledgements

We would like to thank Tony Glover, Paul Luff, Abi Sellen and Jolanda Tromp with whom many of these issues have been discussed and developed. This research has been supported by ESRC Grant No. R000237136. An earlier version of the paper was presented at CSCW'98.

References

1. Barnard, P., May, J. and Salber, D. Deixis and Points of View in Media Spaces: An expirical gesture. *Behaviour and Information Technology*, 15, 1, (1996), 37-50.

2. Barnatt, C. *Cyber Business: Mindsets for a wired age*. Chichester: John Wiley, 1995.

3. Benford, S. D. and Fahlen, L. E. A Spatial Model of Interaction in Virtual Environments, *Proc. ECSCW'93* , 1993, Kluwer Academic Publishers.

4. Benford, S. D., Greenhalgh, C. M. and Lloyd, D. Crowded Collaborative Environments, *Proc. CHI'97* (1997), ACM Press.

5. Bowers, J., Pycock, J. and O'Brien, J. Talk and Embodiment in Collaborative Virtual Environments, in *Proc. CHI'96* (1996), ACM Press.

6. Bowers, J., O'Brien, J. and Pycock, J. Practically Accomplishing Immersion: Cooperation in and for Virtual Environments, in *Proc. CSCW'96* (1996), ACM Press, 380-389.

7. Cruz-Neira, C., Sandin, D. J., DeFant, T. A., Kenyon, R. V. and Hart, J. C., The Cave – Audio Visual Experience Virtual Environment, CACM, 1992, 35 (6), pp 65–72.

8. Furnas, G.W. Generalized Fisheye Views, in *Proc. CHI'86* (April 1986), ACM Press.

9. Garfinkel, H. *Studies in Ethnomethodology*, Polity Press, Cambridge, 1967.

10. Gaver, W.W., Sellen, A., Heath, C.C. and Luff, P. One is not enough: Multiple Views in a Media Space, in *Proc. INTERCHI '93* (April 1993), 335-341.

11. Goffman, E. *Interaction Ritual: Essays on face-to-face behaviour*. New York: Pantheon Books, 1967**.**

12. Goodwin, C. and Goodwin, M.H. Seeing as Situated Activity: Formulating planes. In Engeström, Y. & D. Middleton (eds.) *Communication & Cognition at Work*. Cambridge University Press, Cambridge, 1996, 61-95.

13. Greenhalgh, C.M. and Benford, S.D. Virtual Reality Teleconferencing: Implementation and experience, in *Proc. ECSCW'95* (Stockholm, Sept. 1995).

14. Guye-Vuilleme, A., Capin, T. K., Pandzic, I. S., Thalmann, N. M. and Thalmann, D. Nonverbal Communication Interface for Collaborative Virtual Environments, in *Proc. CVE'98* (June 1998), 105-112.

15. Heath, C. and Hindmarsh, J. Les objets et leur environnement local. La production interactionnelle de réalitiés matérielles. *Raison Pratiques. Cognition et Information en Société*, 8, (1997), 149-176.

16. Heath, C., Luff, P. and Sellen, A. Reconsidering the Virtual Workplace. In Finn, K, Sellen, A. & S. Wilbur (eds.) *Video-Mediated Communication*. Lawrence Erlbaum, New Jersey, 1997.
17. Heath, C. Jirotka, M., Luff, P. and J. Hindmarsh. Unpacking Collaboration: The interactional organisation of trading in a city dealing room. *Computer-Supported Co-operative Work*, *3* (1995), 147-165.
18. Hindmarsh, J. The Interactional Constitution of Objects. Unpublished Ph.D. thesis, U. of Surrey, 1997.
19. Kuzuoka, H., Kosuge, T. and Tanaka, M. GestureCam: A video communication system for sympathetic remote collaboration, in *Proc. CSCW'94* (1994), ACM Press, 35-43.
20. Reynard, G., Benford, S., Greenhalgh, C. and Heath, C. Awareness Driven Video Quality of Service in Collaborative Virtual Environments, in *Proc. CHI'98*, (LA, April 1998), ACM Press.
21. Robertson, G., Czerwinski, M. and van Dantzich, M. Immersion in Desktop Virtual Reality, in *Proc. UIST'97*, (Canada, October 1997), ACM Press.
22. Sacks, H. *Lectures on Conversation* (Vols. 1 & 2), ed. G. Jefferson. Oxford: Blackwell, 1992
23. Schutz, A. *On Phenomenology & Social Relations*. U. of Chicago Press, Chicago, 1970.
24. Smith, R.B., Hixon, R. and Horan, B. Supporting Flexible Roles in a Shared Space, in *Proc. CSCW'98* , (Seattle, November 1998), ACM Press, 197-206.
25. Steed, A., Slater, M., Sadagic, A., Bullock, A. and Tromp, J. Leadership and Collaboration in Shared Virtual Environments, in Proc. VR'99, (Houston, March 1999), IEEE.
26. Tang, J., Isaacs, E. & Rua, M. Supporting Distributed Groups with a Montage of Lightweight Interactions, in *Proc. CSCW'94* (1994), ACM Press, 23-34.