Object-Focused Interaction in Collaborative Virtual Environments

JON HINDMARSH
King's College London
MIKE FRASER
University of Nottingham
CHRISTIAN HEATH
King's College London
and
STEVE BENFORD and CHRIS GREENHALGH
University of Nottingham

This paper explores and evaluates the support for object-focused interaction 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 are: the use of semidistorted views to support peripheral awareness; more explicit or exaggerated representations of actions than are provided by pseudohumanoid avatars; 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 realized.

Categories and Subject Descriptors: H.4.3 [**Information Systems Applications**]: Communications Applications—*Computer conferencing, teleconferencing, and videoconferencing*; H.5.1

This research has been supported by ESRC Grant No. R000237136. An earlier version of the paper was presented at CSCW'98.

Authors' addresses: J. Hindmarsh, Work, Interaction & Technology Research Group, The Management Centre, King's College London, Franklin Wilkins Building, 150 Stamford Street, London, SE1 8WA, UK; email: jon.hindmarsh@kcl.ac.uk; M. Fraser, Communications Research Group, School of Computer Science and IT, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, UK; email: mcf@cs.nott.ac.uk; C. Heath, Work, Interaction & Technology Research Group, The Management Centre, King's College London, Franklin Wilkins Building, 150 Stamford Street, London, SE1 8WA, UK; email: christian.heath@kcl.ac.uk; S. Benford and C. Greenhalgh, Communications Research Group, School of Computer Science and IT, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, UK; email: sdb@cs.nott.ac.uk; cmg@cs.nott.ac.uk.

[**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artificial, augmented and virtual realities*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Evaluation/methodology*

General Terms: Human Factors, Experimentation, Design

Additional Key Words and Phrases: Social interaction, virtual reality, CSCW, objects, shared spaces, user interface design, embodiment

1. INTRODUCTION

Recent years have witnessed extraordinary advances in the quality and effectiveness of visualization and VR technologies. Indeed, a wide variety of organizations 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 visualizations 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 recognized that changes in the structure of contemporary organizations will place corresponding demands on technology to provide support for such distributed collaborative work [Barnatt 1995]. The trend toward disaggregation, globalization, and dynamic networks of firms suggests that CSCW technologies will play an increasingly important part in supporting cooperation 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. For example, it is widely recognized that much collaborative work rests upon the sharing and discussion of a whole host of documents, tools, and other artefacts and that supporting such interaction is a general problem for the development of advanced shared workspaces. Although asynchronous text-based systems to support remote work, such as email, Notes, and the World Wide Web, are flourishing within the business community, technologies to support real-time, collaborative work, such as media spaces, 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 a variety of workplace objects (e.g., Heath et al. [1997]). This would seem of critical importance for the development of virtual environments where remote colleagues would need to discuss virtual objects and scenes as well as collaboratively coordinate their navigation around, and looking within, the world. Thus, we need to draw on our understanding of the kinds of resources people utilize in face-to-face interaction to conduct object-focused work, to explore the problems and difficulties faced by individuals interacting through CVEs.

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.

2. BACKGROUND

In their wide-ranging investigation of organizational 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., Goodwin and Goodwin [1996], Heath and Hindmarsh [2000], and Heath et al. [1994]). 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 coordinate those activities, in real time, with the conduct of others. Indeed, it is found that many activities within colocated 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 [Hindmarsh and Heath 2000]. 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., Tang et al. [1994], Kuzuoka et al. [1994], Gaver et al. [1993], and Heath et al. [1997]). However, it is not clear that such systems provide adequate support for object-focused collaboration.

One of the author's earlier attempts to develop a media space to support variable access between participants and their respective domains also met with limited success [Gaver et al. 1993; Heath et al. 1997]. In MTV II, an experiment undertaken with Bill Gaver, Abi Sellen, and Paul Luff, Heath 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 coparticipant was referring to, and, more specifically, where, and at what, they were 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 [Heath and Hindmarsh 2000]. This undermined even simple collaboration in and through the mutually available objects and artefacts (for similar findings, see Barnard et al. [1996]).

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 endeavor 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, e.g., in the form of embedded video views that are displayed as dynamic texture maps attached to virtual objects [Reynard et al. 1998]. 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 recognized, it is often argued that problems of recognizing what views and scenes are available to the other are "naturally" overcome in 3D worlds (e.g., Smith et al. [1998]). 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 (or avatars) 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 organization of human conduct in collaborative virtual environments. Bowers, Pycock, and O'Brien have begun to explore such issues through two observational analyses of research meetings carried out between participants located at five sites and in three countries (UK, Sweden, and Germany). One of their studies [Bowers et al. 1996a] describes the interactional use of simple embodiments in virtual collaboration by showing how changes in the orientation and movement of the embodiments is coordinated with emerging features of the talk. These authors have also discussed the ways in which actions in a virtual world, especially those that seem somehow odd, can be related to activities within the users' real-world environment [Bowers et al. 1996b], that is to say, how ongoing activities and interactions involving a participant in their physical environment can impact upon the ways in which their actions appear and are treated in the virtual world. More recently, Steed et al. conducted experiments in small-group behavior [Steed et al. 1999]. This work utilizes statistical analysis to suggest positive relationships between copresence and presence, presence and immersion, presence and group accord, and argues that immersive-style interfaces may confer leadership status to a participant during collaborative tasks. However, such studies are rare, and interestingly, no research has explored the ways in which the visible properties of the virtual environment (i.e., objects other than the embodiments) feature in, and are related to the intelligibility of, the participants' actions and activities.

3. EXPERIMENTING WITH FURNITURE WORLD

To investigate object-focused interaction in CVEs, we adapted a task from the previous studies of the MTV system [Gaver et al. 1993]. 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 [Benford et al. 1997].

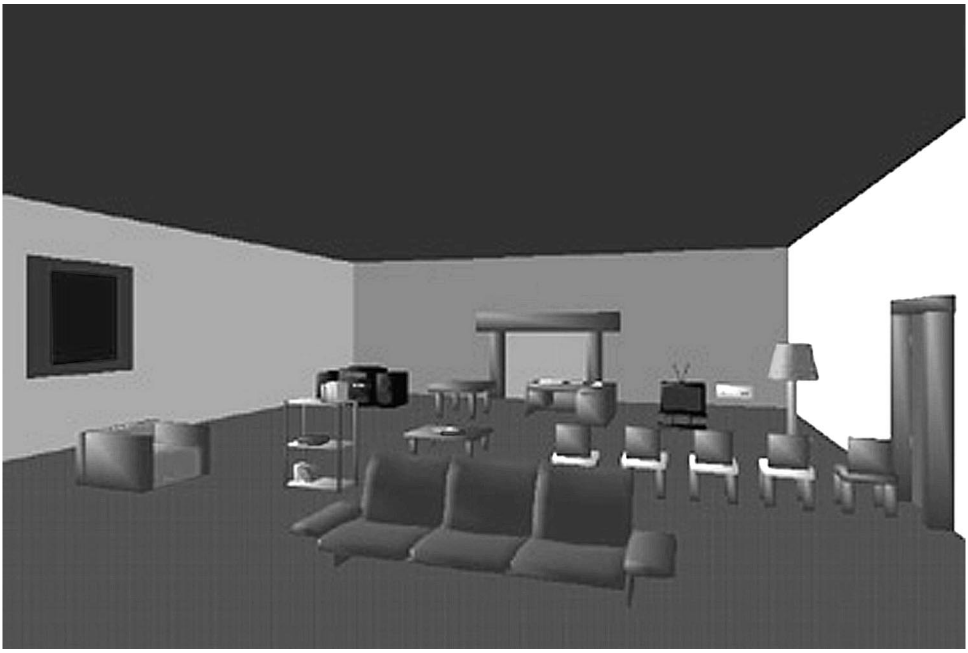


Fig. 1. An overview of furniture world.

MASSIVE-2 allows multiple participants in a shared virtual world to communicate using a combination of 3D graphics and real-time audio. Participants are also able to grasp and move virtual objects in the world. 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 encourages (or even demands) 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.

3.1 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.

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,

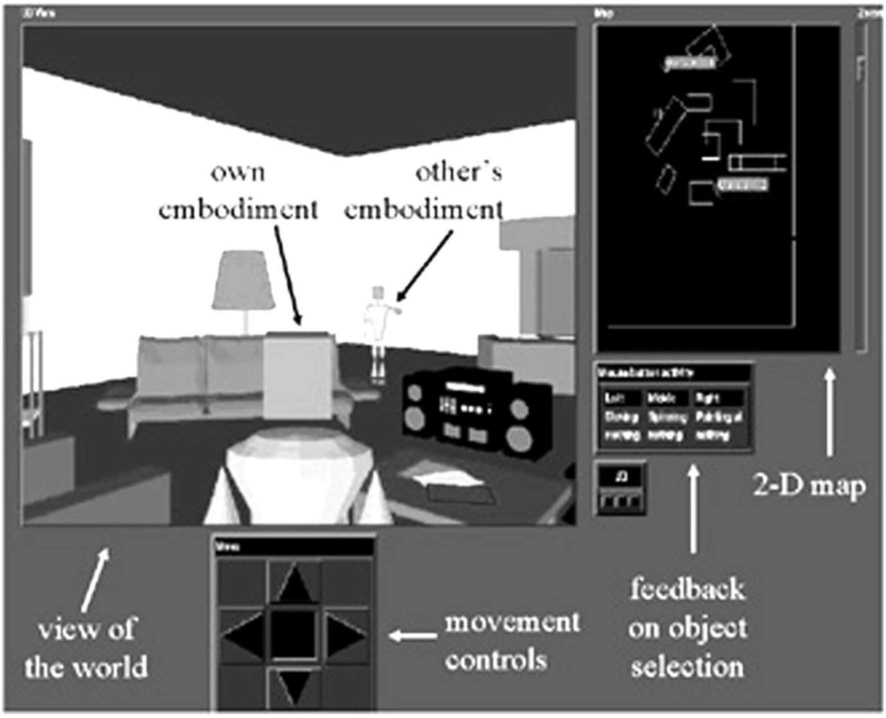


Fig. 2. The furniture world interface.

this meant that the embodiment should be generally humanoid (i.e., have a recognizable 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 pseudohumanoid 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 toward it, as shown in Figure 3.

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 toward 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

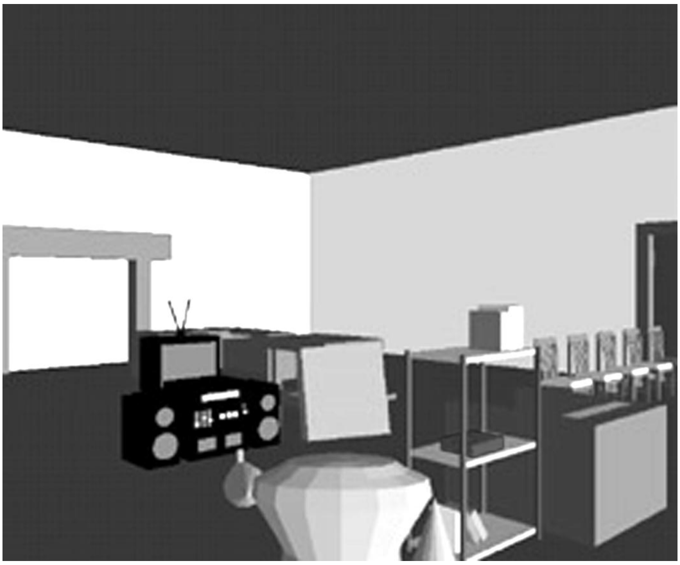


Fig. 3. A user points at the stereo—the embodiment's "head" and "arm" orient toward the selected object.

(i.e., 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.

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 [Greenhalgh and Benford 1995]. 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 minimize distortion on the desktop interface.

3.2 Data Collection

Six trials of two participants and two trials of three participants were performed. Most participants were students, 12 males and six females, with a broad mixture of previous acquaintance. None of them had a

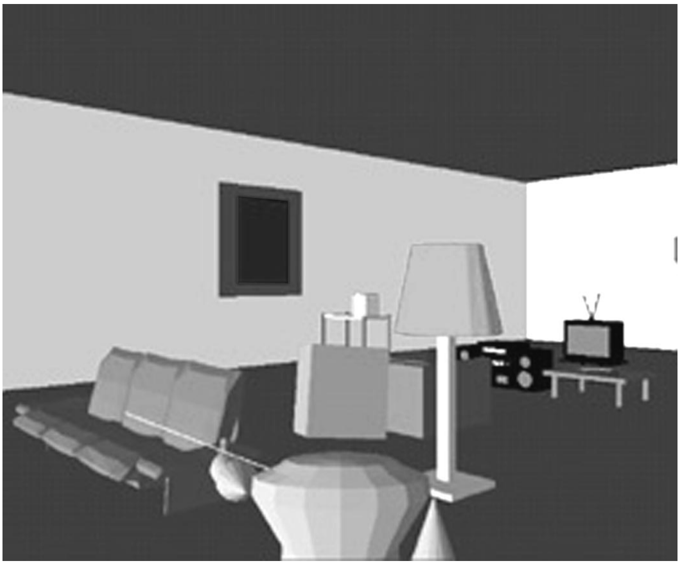


Fig. 4. A user grasps the sofa—the embodiment's head and arm orient toward the grasped object, and a line extends out from the avatar to the sofa.

background in CVE technology. Each trial took about an hour and consisted of 10 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 Figure 5) and contained audio from the participant in the shot.

The analysis of these recordings draws on conversation analysis [Sacks 1992] 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 organization of interaction in everyday workplaces [Goodwin and Goodwin 1996; Heath et al. 1994] and indeed in multiuser VR [Bowers et al. 1996a; 1996b]. We aim to build on, and contribute to, this work.

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 organization of interaction "through" this communication medium, and therefore it would be extremely premature

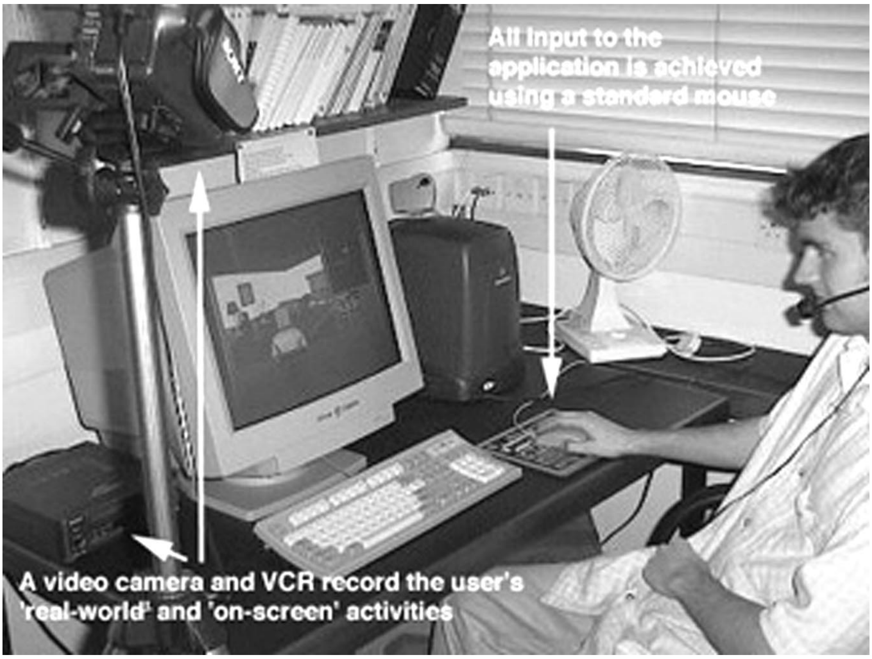

All input to the application is achieved using a standard mouse
A video camera and VCR record the user's 'real-world' and 'on-screen' activities

Fig. 5. A user being filmed.

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 sensitize ourselves to the key issues that impact upon and engender the ways in which individuals interact and discuss objects in CVEs.

4. 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 this and related systems:

—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 degrees).

—Participants compensate for the "fragmenting" of the workspace by using talk to make explicit actions and visual conduct that are recurrently implicit in copresent 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.

We found that these three issues generated problems for all of the participants at various times within the duration of their collaboration. Moreover, they lead to the disruption of, what we have argued, is one of the critical and foundational elements of collaborative work—i.e., the reference to, and discussion of, objects and artefacts.

4.1 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 toward an object. This enables participants to use their virtual embodiments to visibly 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[1] shows the view that Sarah has of Karen's embodiment, her gesture, and the table.

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.

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 [Heath and Hindmarsh 2000]. 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

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

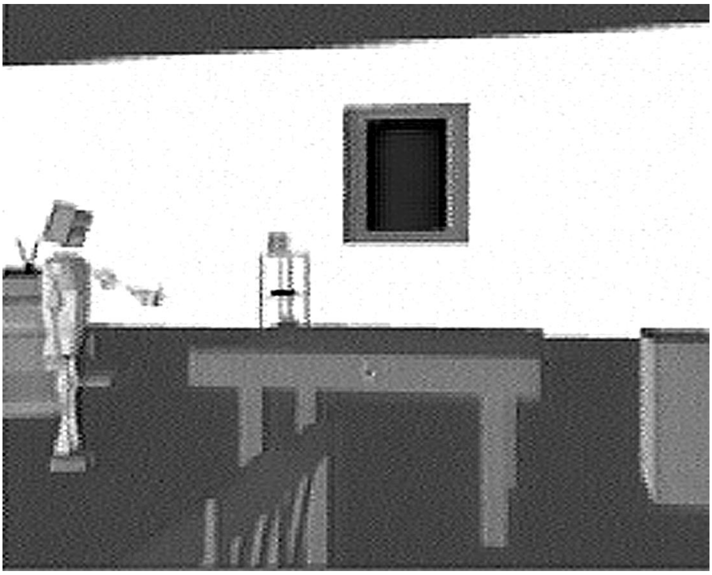


Fig. 6. Sarah's view of Karen's embodiment pointing at the table.

find it problematic to reassemble 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—in this regard, Example 1 provides an exceptional instance. It turns out that it is critical that they do see them in relation to one another. So, they first 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 toward 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 360 degrees, via Sarah's gesture, to return to the very same desk.

In copresent interaction, when an individual asks a coparticipant to look at an object at which they are pointing, that coparticipant can usually see them in relation to their surroundings. They simply turn from the body of the other to find the referenced object [Hindmarsh and Heath 2000]. 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 coparticipant'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 even 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 toward 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) and thus cannot see the objects

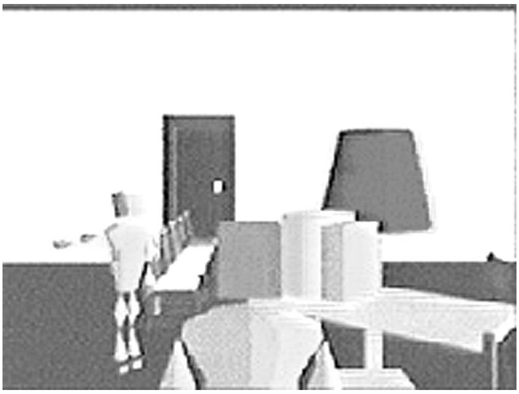


Fig. 7. Andre's view of Trevor's embodiment.

toward which Trevor seems to be pointing, so he asks "what you pointing at?"

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 recognizes that it could be designed for him. Unfortunately, Trevor is just repositioning the television, 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 copresent 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 [Heath and Hindmarsh 2000]. 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.

4.2 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 they are acting. Therefore they use their talk to make explicit certain actions and visual conduct.

For example, in copresent interaction an individual may simply say "what do you know about this?" alongside a gesture. Their coparticipant 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 [Hindmarsh and Heath 2000], i.e., 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 toward an object or location can be much longer than the simple query and response that follows.

Unfortunately, the additional time involved in establishing mutual orientation is not the critical concern. This prefatory sequence actually disrupts the flow and organization of collaborative activities. In copresent 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

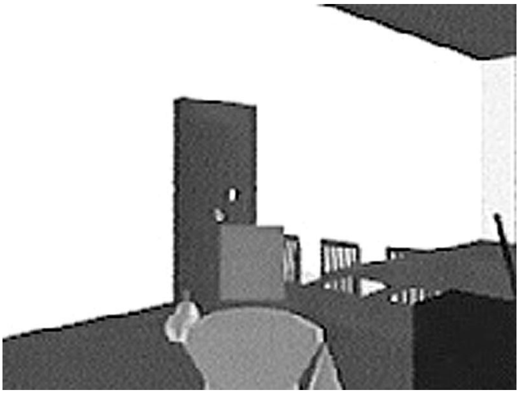


Fig. 8. Trevor's view as he points to the door—note that Andre's embodiment is not visible.

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 copresent 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 coparticipant as they point something out to them. To compensate, their coparticipants 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 toward the door))
A: I'm coming ((A rotates))
T: Hang on ((T repositions gesture))
A: Yeah, okay, I got the door.
```

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 copresent interaction, they are even able to transform the course of a pointing gesture with regard to the emerging orientation of their coparticipant(s) [Hindmarsh and Heath 2000].

In this CVE, the problem for participants is that when they point to something, they often cannot see their coparticipant(s). In copresent interaction, participants routinely configure a pointing gesture and then turn to view their coparticipant's "response" [Hindmarsh and Heath 2000]. 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 organization and coordination of much copresent 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 coordinating and organizing 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.

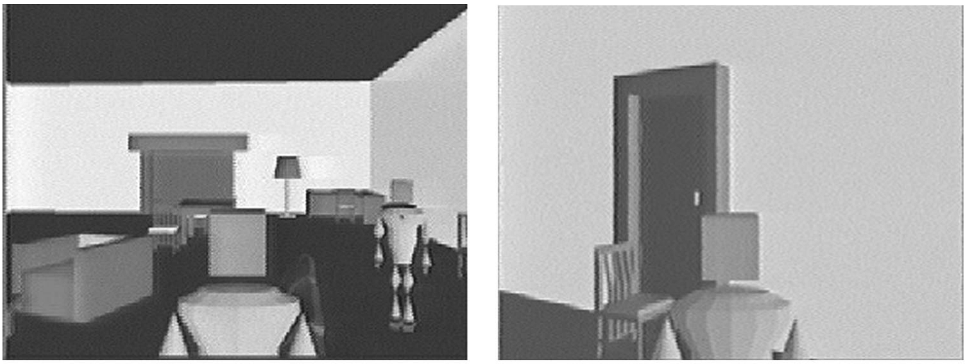
4.3 Hidden Perspectives

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.
```



Pete's View

Rick's View

Fig. 9. A comparison of views.

Although Rick has the fireplace in his view, he does not recognize 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.[2] Compare their viewpoints in Figure 9.

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 pseudohumanoid form is confusing. This kind of embodiment may give participants a sense that it possesses a “human-like” field of view, i.e., 180 degrees. However, the users’ field of view in this CVE is only 55 degrees. 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 coordinate 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, i.e., when one participant moves an object and the other directs that movement. When participants move an object in the

[2] 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.

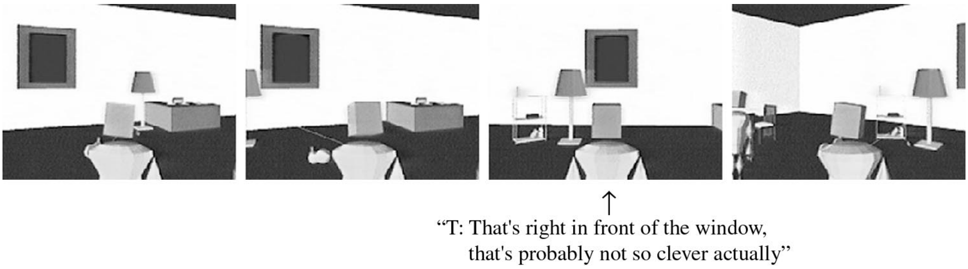

"T: That's right in front of the window,
that's probably not so clever actually"

Fig. 10. Mark's view as he moves the lamp.

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 suggest 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.

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 demeanor. 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 problematizes 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 coparticipants, as they have little sense of how they are engaged in, or orienting to, the ongoing activity. The orientations of the others 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.

5. PRINCIPAL ISSUES

The CVE undermines certain resources commonly used to organize and coordinate collaborative work. Although the system does not prevent task completion 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 copresent 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 coordinate 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 copresent interaction, Schutz suggested that we assume our different perspectives on the world and on an object are irrelevant for the activity at hand [Schutz 1970]. 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 [Heath et al. 1994; Heath and Hindmarsh 2000]. 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 organization, it is suggested that these issues are important for the design of systems to support synchronous remote working.

6. 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:

(1) 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.

(2) 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.

(3) Clumsy movement—movement in CVEs may be slow due to problems of locating destinations, controlling the interface, and system performance.

(4) 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 to provide enhanced support for object-focused cooperative work. This discussion will include possible developments of our system, but the reanalysis and evaluation of those developments is beyond the scope of this paper. Nevertheless, such evaluation work will be undertaken in the course of future research.

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 synchronized back-projected views of a virtual world [Cruz-Neira et al. 1992]. 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 recognizing 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.

6.1 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 is important, as indeed is 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 accurate 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

[Greenhalgh and Benford 1995]. For example, locating the camera behind their embodiment and 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—for example, work on the MTV-1 system revealed that the feedback monitor was not intuitively used to assist object-focused discussions [Gaver et al. 1993; Heath et al. 1997].

Another possibility is to introduce more controlled perspective distortions than would be obtained by just widening the field of view in software. Techniques for utilizing perspective distortions such as fish-eye lenses [Furnas 1986] could be applied for interfaces to virtual environments. More recently, Robertson et al. introduced the related approach of peripheral lenses [Robertson et al. 1997]. 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. 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-degree field of view utilizing heavy perspective distortion. We allocate 8 inches for the main view, and this also renders a 60-degree field of view, but with very little distortion (see Figure 11).

Our implementation also extends peripheral lenses with the ability to focus on either periphery, through a technique we coin "peripheral glancing." Interestingly, Pierce et al. address the limited capabilities in single-user 3D systems for accessing tools and information on one's own avatar through a "glancing" metaphor [Pierce et al. 1999]. They describe the limitation in both field of view and viewpoint manipulation techniques, and whilst not providing explicit peripheral visual information, they note that glances may be a useful technique because "by combining glances and more

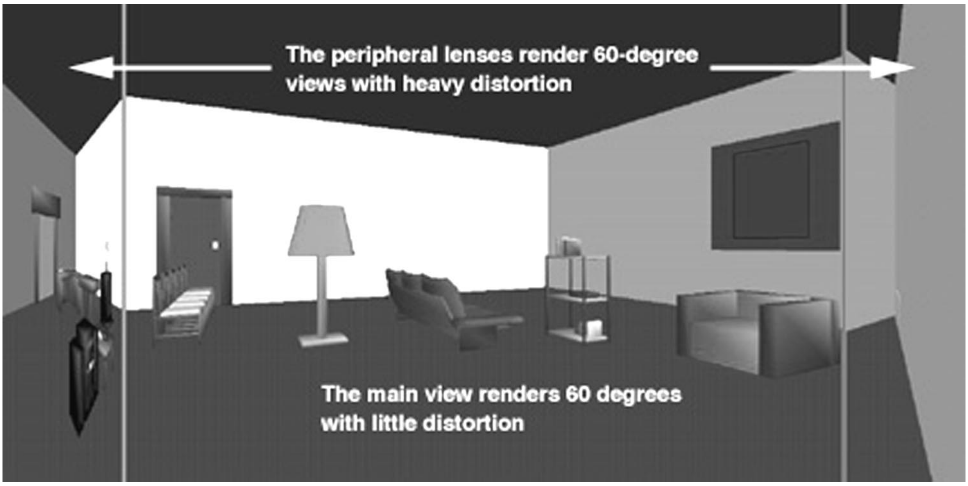

The peripheral lenses render 60-degree views with heavy distortion
The main view renders 60 degrees with little distortion

Fig. 11. View of furniture world using peripheral lenses.

traditional 3D navigation, we create a new metaphor for viewpoint control that includes both a virtual head and virtual body."

In our system, 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.

6.2 Representation of Actions and Pseudorealism in Embodiment

Our second proposal focuses on the issue of providing better information about others' actions. Developers of 3D 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 experience difficulties in finding the object being discussed or referred to. Thus, we propose that the pseudorealistic 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 the 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.

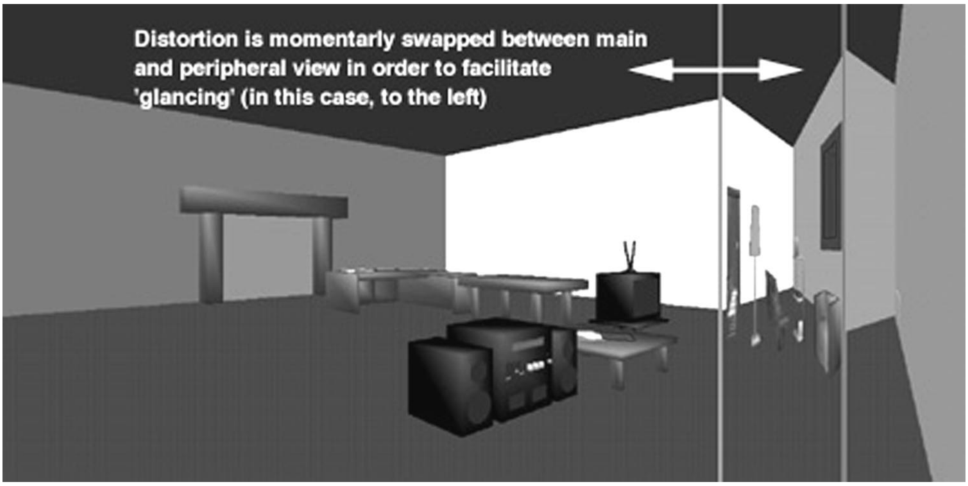

Distortion is momentarly swapped between main and peripheral view in order to facilitate 'glancing' (in this case, to the left)

Fig. 12. Peripheral glancing in furniture world.

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 labeling 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 also 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.

In general, we propose that the representation of actions within CVEs should be designed through a consideration of whether and how each action is represented on the source embodiment, the target object(s), and in the intervening environment. Furthermore, these representations should be both consistent and distinguishable. For example, in our system, 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.

Thus, we propose extending the representation 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 embodiment's view (the extent of its field of view) visible as a semitransparent frustrum. This extension is shown in Figure 13 where the other participant's view is 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 whether other techniques are more effective.

We can also extend the representation of pointing to include target objects and the surrounding environment. The former involves highlighting

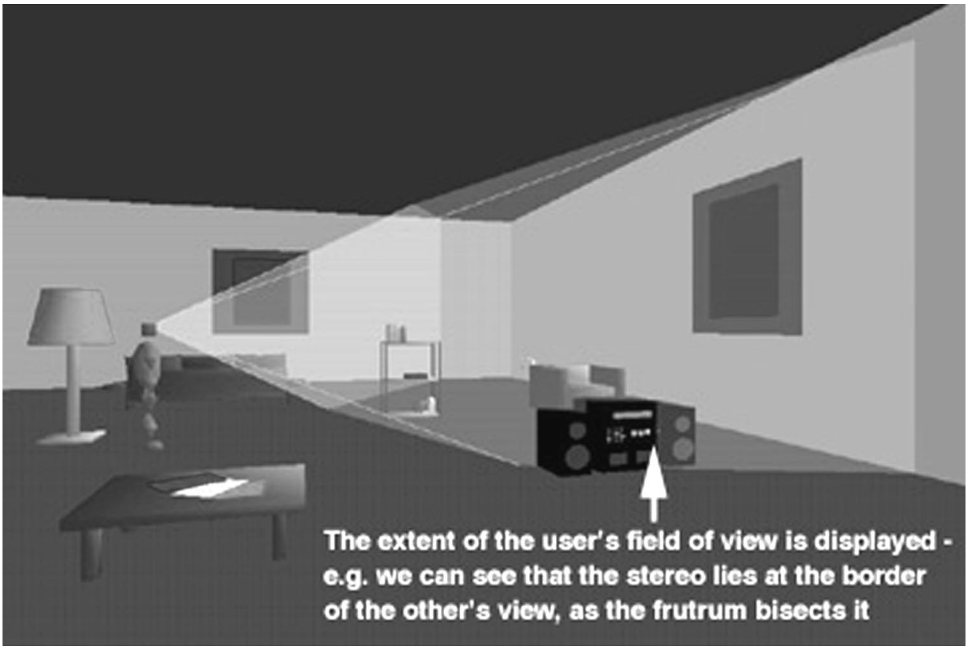

The extent of the user's field of view is displayed -
e.g. we can see that the stereo lies at the border
of the other's view, as the frutrum bisects it

Fig. 13. Visible view frustrum in furniture world.

the targets. The latter involves rendering a visible ray of light between the embodiment and this target or enabling users to extend an arm to point at distant objects and scenes.

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 it from pointing. Figure 14 shows an example of representing grasping in this way (it is taken from the perspective of the grasping embodiment).

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., Guye-Vuilleme et al. [1999]). Indeed, some may argue that the embodiments used in this experiment were not realistic

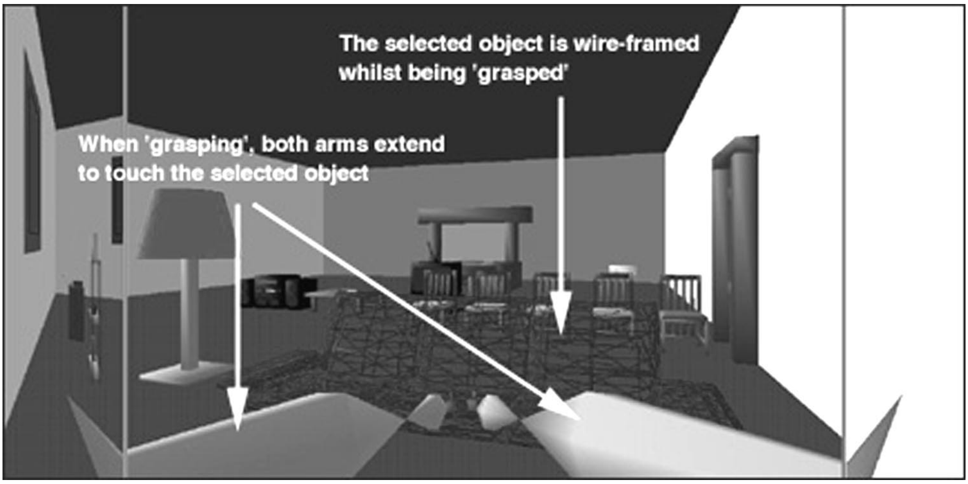

The selected object is wire-framed whilst being 'grasped'
When 'grasping', both arms extend to touch the selected object

Fig. 14. Grasping and wire-framing an object with elongated arms.

enough and that this was the source of users' problems. However the findings would suggest that the straightforward 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 behaviors 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.

6.3 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 toward relevant targets, as a way of dealing with clumsy navigation in CVEs. Furthermore, in a cooperative 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 prioritizing 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 could 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" or distracted from their involvement in the task and interaction at hand, by virtue of their need to concentrate on switching between, and selecting, different views.

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., who discuss experiments with the GestureCam system in which instructors commented on tasks by remote participants [Kuzuoka et al. 1994]. They had a camera mounted on a small robot, 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 to get a better view of some object, "the instructor often lost track of his position" [Kuzuoka et al. 1994, p. 40]. The rapid change in position of the camera disoriented the instructor and disrupted his geography of the remote space. 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 coparticipants. Furthermore, we decided not to simply "snap" the view to that of the other, because, as the empirical sections imply, it is often essential that the objects under discussion are seen in relation to the other's avatar. Thus, we aim to preserve the fluidity of the cooperative interaction and maintain the users' overall sense and knowledge of the space, whilst providing participants with the resources to "connect" and understand actions in relation to relevant objects.

More generally, we suggest that support for navigation, and coparticipation, 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 the 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 [Benford and Fahlen 1997] 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.

6.4 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.

6.5 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 do, 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 organization 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. The system could be designed to be sensitive to the activity underway and respond to changes between activities. 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.

7. 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 pseudohumanoid 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 cooperative 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, e.g., it should be easy to turn toward an object that someone else is pointing at or grasping.

Although the evaluation of these redesign proposals is beyond the scope of this paper, it is very much an activity for immediate future work. In the near future we will conduct a repeat of these experiments using a next generation of the system and interface, in line with the proposals made here. Indeed, we have already explored some of the proposals in related work (see Fraser et al. [1999]). However, we feel it is critical that we undertake a full and thorough comparative evaluation of the various redesign proposals within the context of the original furniture world task before setting them challenges that will inevitably arise when exposed to alternative activities and application domains and when we increase the numbers of participants. Indeed, it is to be hoped that the unanticipated consequences (both positive and negative) of the redesigns are as interesting as the consequences of the original CVE design.

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" [Garfinkel 1967]. They dramatically reveal presuppositions and resources which often remain unexplicated in more conventional studies of the workplace. Whilst we believe these presuppositions 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.

ACKNOWLEDGMENTS

We would like to thank Paul Luff, Abi Sellen, and Jolanda Tromp with whom many of these issues have been discussed and developed. Our thanks also to the anonymous reviewers for their constructive comments on the earlier version of this paper.

REFERENCES

BARNARD, P., MAY, J., AND SALBER, D. 1996. Deixis and points of view in media spaces: An empirical gesture. *Behav. Inf. Tech. 15*, 1, 37–50.

BARNATT, C. 1995. *Cyber Business: Mindsets for a Wired Age*. John Wiley and Sons Ltd., Chichester, UK.

BENFORD, S. D. AND FAHLEN, L. E. 1993. A spatial model of interaction in virtual environments. In *Proceedings of the 3rd European Conference on Computer-Supported Cooperative Work* (ECSCW '93). Kluwer B.V., Deventer, The Netherlands, 109–124.

BENFORD, S., GREENHALGH, C., AND LLOYD, D. 1997. Crowded collaborative virtual environments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '97, Atlanta, GA, Mar. 22–27), S. Pemberton, Ed. ACM Press, New York, NY, 59–66.

BOWERS, J., PYCOCK, J., AND O'BRIEN, J. 1996a. Talk and embodiment in collaborative virtual environments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '96, Vancouver, B.C., Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 58–65.

Bowers, J., O'Brien, J., and Pycock, J. 1996b. Practically accomplishing immersion: Cooperation in and for virtual environments. In *Proceedings of the 1996 ACM Conference on Computer-Supported Cooperative Work* (CSCW '96, Boston, MA, Nov. 16–20), M. S. Ackerman, Ed. ACM Press, New York, NY, 380–389.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. 1992. The CAVE: Audio visual experience automatic virtual environment. *Commun. ACM 35*, 6 (June), 64–72.

Fraser, M., Benford, S., Hindmarsh, J., and Heath, C. 1999. Supporting awareness and interaction through collaborative virtual interfaces. In *Proceedings of the ACM Symposium on User Interface Software Technology* (UIST '99, Asheville, NC, Nov.). ACM, New York, NY.

Furnas, G. W. 1986. Generalized fisheye views. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '86, Boston, MA, Apr. 13–17), M. Mantei and P. Orbeton, Eds. ACM Press, New York, NY.

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

Gaver, W. W., Sellen, A., Heath, C., and Luff, P. 1993. One is not enough: Multiple views in a media space. In *Proceedings of the ACM Conference on Human Factors in Computing* (INTERCHI '93, Amsterdam, The Netherlands, Apr. 24–29), B. Arnold, G. van der Veer, and T. White, Chairs. ACM Press, New York, NY, 335–341.

Goodwin, C. and Goodwin, M. H. 1996. Formulating planes: Seeing as situated activity. In *Cognition and Communication at Work*, Y. Engeström and D. Middleton, Eds. Cambridge University Press, New York, NY.

Greenhalgh, C. M. and Benford, S. D. 1995. Virtual reality teleconferencing: Implementation and experience. In *Proceedings of the European Conference on Computer-Supported Cooperative Work* (ECSCW '95, Stockholm, Sweden, Sept.). Kluwer B.V., Deventer, The Netherlands.

Guye-Vuilleme, A., Capin, T. K., Pandzic, I. S., Thalmann, N. M., and Thalmann, D. 1999. Nonverbal communication interface for collaborative virtual environments. *Virt. Real. 4*, 1, 49–59.

Heath, C. and Hindmarsh, J. 2000. Configuring action in objects: From mutual space to media space. *Mind. Cult. Act. 7*, 1/2, 81–104.

Heath, C., Luff, P., and Sellen, A. 1997. Reconsidering the virtual workplace. In *Video-Mediated Communication*, K. E. Finn, A. J. Sellen, and S. B. Wilbur. Lawrence Erlbaum Associates Inc., Hillsdale, NJ, 323–349.

Heath, C., Jirotka, M., Luff, P., and Hindmarsh, J. 1994. Unpacking collaboration: The interactional organisation of trading in a city dealing room. *Comput. Supp. Coop. Work 3*, 2, 147–165.

Hindmarsh, J. and Heath, C. 2000. Embodied reference: A study of deixis in workplace interaction. *J. Prag. 32*, 12, 1855–1878.

Kuzuoka, H., Kosuge, T., and Tanaka, M. 1994. GestureCam: A video communication system for sympathetic remote collaboration. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (CSCW '94, Chapel Hill, NC, Oct. 22–26), J. B. Smith, F. D. Smith, and T. W. Malone, Chairs. ACM Press, New York, NY, 35–43.

Pierce, J. S., Conway, M., Van Dantich, M., and Robertson, G. 1999. Toolspaces and glances: Storing, accessing, and retrieving objects in 3D desktop applications. In *Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics* (Apr.), J. Hodgins and J. D. Foley, Eds. ACM Press, New York, NY.

Reynard, G., Benford, S., Greenhalgh, C., and Heath, C. 1998. Awareness driven video quality of service in collaborative virtual environments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '98, Los Angeles, CA, Apr. 18–23), C.-M. Karat, A. Lund, J. Coutaz, and J. Karat, Eds. ACM Press/Addison-Wesley Publ. Co., New York, NY, 464–471.

Robertson, G., Czerwinski, M., and van Dantzich, M. 1997. Immersion in desktop virtual reality. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology* (UIST '97, Banff, Alberta, Canada, Oct. 14–17), G. Robertson and C. Schmandt, Chairs. ACM Press, New York, NY, 11–19.

SACKS, H. 1996. *Lectures on Conversation*. Blackwell Publishers, Inc., Cambridge, MA.

SCHUTZ, A. 1970. *On Phenomenology & Social Relations*. University of Chicago Press, Chicago, IL.

SMITH, R. B., HIXON, R., AND HORAN, B. 1998. Supporting flexible roles in a shared space. In *Proceedings of the 1998 ACM Conference on Computer-Supported Cooperative Work* (CSCW '98, Seattle, WA, Nov. 14–18), S. Poltrock and J. Grudin, Chairs. ACM Press, New York, NY, 197–206.

STEED, A., SLATER, M., SADAGIC, A., BULLOCK, A., AND TROMP, J. 1999. Leadership and collaboration in shared virtual environments. In *Proceedings of the Conference on VR'99* (VR'99, Houston, TX, Mar.). IEEE Press, Piscataway, NJ.

TANG, J. C., ISAACS, E. A., AND RUA, M. 1994. Supporting distributed groups with a montage of lightweight interactions. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (CSCW '94, Chapel Hill, NC, Oct. 22–26), J. B. Smith, F. D. Smith, and T. W. Malone, Chairs. ACM Press, New York, NY, 23–34.