

# Hands on Hitchcock: Embodied Reference to a Moving Scene

Paul Luff<sup>1</sup>, Naomi Yamashita<sup>2</sup>, Hideaki Kuzuoka<sup>3</sup>, Christian Heath<sup>1</sup>

<sup>1</sup> King's College, London  
150 Stamford Street, London,  
UK

<Paul.Luff@kcl.ac.uk,  
Christian.Heath@kcl.ac.uk

<sup>2</sup> NTT Communication Science  
Laboratories,  
2-4 Hikaridai Seika-cho  
Soraku-gun, Kyoto, Japan  
<naomiy@acm.org>

<sup>3</sup> University of Tsukuba, Japan,  
1-1-1 Tennoudai  
Tsukuba, Ibaraki, Japan  
<kuzuoka@iit.tsukuba.ac.jp>

## ABSTRACT

In this paper we report on some experiments with a high fidelity media space, t-Room, an immersive system that presents full scale, real-time images of co-participants who are in similar spaces many miles apart. Although being designed to provide a coherent environment for interaction the system introduces a number of incongruities, both in time and space. Drawing on some quasi-naturalistic experiments, where the participants were required to analyse complex data, we consider how the participants manage these incongruities. We conclude by briefly discussing the resources people utilize to produce and recognize conduct in embodied spaces.

## Author Keywords

Media spaces, interaction analysis, CSCW, embodied interaction.

## ACM Classification Keywords

H5.3. Information interfaces and presentation (e.g., HCI): Computer-supported cooperative work. H4.3.

## General Terms

Human Factors.

## INTRODUCTION

Of the early efforts to develop systems to support collaborative work perhaps the most notable were attempts to design media spaces. In the 1980s and 1990s leading research laboratories and university departments introduced audio-visual infrastructures and networked computers to explore the potential for technologies to support everyday work activities in real-time. Typically these efforts were associated with office work between two people in two different locations [e.g. 10, 25]. Analysis of the conduct of participants in these spaces suggested that there were problems referring to features of the environment, notably paper documents; that the

shapes of gestures, even simple pointing, were transformed when seen and produced through video; and that the technologies introduced a number of asymmetries into the interactional environment in which actions were performed [16]. Although a number of studies reported the advantages of long-term open video connections [4], these infrastructures tended not to persist for longer than the research projects that introduced them.

Recently there has been a renewed interest in media spaces [15]. On the one hand, higher broadband, high definition and high-fidelity systems like HP's Halo and Cisco's Telepresence offer capabilities that seem to overcome many of the problems found in early media spaces. For activities like meetings, what have become to be known as blended spaces [1], have been carefully designed to provide a coherent environment for interaction. On the other, new applications have emerged for high-quality, video-mediated environments. In the fields of e-science and cyber-infrastructure, technologies like the AccessGrid provide sophisticated support for scientists to collaborate when they are remote, either for distributed research meetings or when they are engaged in activities between laboratories [5]. And yet, it is unclear the extent to which these novel environments do resolve the interactional problems found in the early media spaces. There are few, if any, detailed studies of their use and despite providing resources to, say make a document available at both sites, such systems provide little support for integrating the use of other artefacts within the interaction. There are still ways in which the spaces are disjoint, so that even a simple action like pointing to an object in a remote domain can be problematic.

In this paper, we examine the use of one particular high definition, high fidelity system, t-Room. This is an experimental, immersive environment that allows participants to share and collaborate over remote electronically presented materials. However, rather than just the hands of two participants being projected in the remote domain [cf. 8, 21, 24] full 'embodiments' of several parties are visible. Participants can move around a small office-like space and change how they arrange themselves in the environment and this is reflected in a similar space at a remote site. The technology is designed

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Copyright 2011 ACM 978-1-4503-0267-8/11/05...\$10.00.

to offer a symmetrical environment in which distributed actions are accomplished. We report on some quasi-naturalistic experiments that were developed to investigate the nature of multi-party embodied interaction in this innovative space. They reveal that even a sophisticated space like t-Room subtly transforms how activities are displayed in a remote domain. We reveal how participants manage these transformations and discuss the implications of this analysis for the design of high-definition, high fidelity, video-based systems to support collaboration, our understanding of embodied interaction, and how we might support complex artefact-based interaction and collaboration with technology.

## BACKGROUND

From the earliest media spaces to the most recent high fidelity distributed meeting spaces a common problem facing designers and users is how to integrate documents, whether these are paper or electronic, into the mediated environment. In early media spaces, even if augmented with a document camera, there were difficulties establishing a mutual orientation to details on a piece of paper [16]. It is unclear if these problems are resolved by more recent attempts at video-mediated interaction. It is not just a problem deciding where to display the document, but also how to present the conduct of the co-participants in and around the document. This may be due to assumptions about the activities these systems were designed to support. Early media spaces focused on supporting face-to-face interaction. Therefore, it is quite understandable that activities involving the use of artefacts were neglected. More recent innovations in telepresence have focused on 'business meetings' where any documents, such as PowerPoint slides, are typically displayed on monitors in fixed positions either above or below the images of the remote participants. Apart from the use of a computer cursor there are few resources to identify and refer to details of documents.

Such capabilities are important if richer forms of collaboration are required. Recent developments in e-research (or cyberinfrastructures) have drawn upon video mediated communications technologies provided by the AccessGrid to support scientists, social scientists and other researchers to collaborate in real-time over common materials [2, 7, 30]. Here again, there has been a challenge to integrate documents into these infrastructures. It is not just that the materials are visually complex, but researchers need to be able to refer to features in complex ways, not just pointing but also animating the images through their gestures and visual conduct [6, 20].

Efforts in CSCW and elsewhere to provide richer environments for interaction have revealed the difficulties of designing ways to support referential practices. Providing avatars in Collaborative Virtual Environments (CVEs) with ways of pointing to objects in the shared space fragmented the ecology in which gestures are

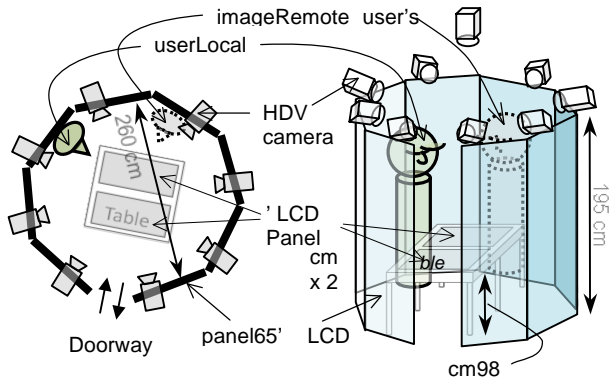
produced and recognized [17]. Similar efforts of augmenting video-mediated interaction fractured the environment of interaction. Even with the addition of robots with pointing arms equipped with laser pointers, participants found it hard to tie the activities of the person pointing with the object being pointed at [23]. Some of these issues do seem to be resolved by systems that capture and present richer kinds of visual conduct [e.g. 18, 19]. By sophisticated video capture and projection techniques these have aimed to integrate real-time presentations of the hands of remote colleagues into the workspace and seem effective for supporting collaboration over detailed documents between two participants. Given the nature of the technology it is hard to envisage how such techniques could be scaled up when there is more than one party at each site.

One critical problem that has persisted through the early developments in media spaces, high fidelity conferencing systems and perhaps for most advanced collaborative technologies is that of reference. In video-mediated environments this could be glossed as the problem of presenting and understanding pointing. At first, this may appear a simple problem – the solution needing only for the person doing the pointing to be able to accurately identify the object being referred to. However, this neglects the rich variety of ways in which pointing can be achieved, as revealed by observing the work of the hands [21, 24]. The solution may then to be provide accurate representations of the hands, what could be considered as disembodied embodiments. Although providing greater fidelity such solutions can easily neglect how pointing is accomplished collaboratively [14]. Pointing and other referential gestures are produced in the light of the conduct of a c-participant, the 'pointer' monitoring how the other, the recipient or 'the pointee', (as Tallis describes them in a recent work on pointing) is, in the course of the pointing, making sense of the ongoing gesture [28]. Indeed, Tallis has not only argued that the complexity of this seemingly simple action reveals the distinctive nature of human collective consciousness, but also the ways in which humans are unique. Notwithstanding the grandiloquence of these claims it is perhaps understandable why trying to reproduce collaborative referential conduct has proved problematic for designers of co-operative systems.

In this paper, we will consider another attempt at developing a system, called t-Room, to provide a rich environment for collaboration. By presenting real-time, full size, embodiments of remote co-participants, t-Room can offer ways for those participants to produce a wide range of 'naturalistic' pointing actions, but also to produce these in the light of the conduct of their remote colleagues. Also, the scale of the system allows the participation of more than two parties and hence provides a resource for investigating the accomplishment of multi-party pointing.

## IMMERSIVE MEDIA SPACES

T-Room is an experimental room-sized video conferencing system that aims to support group-to-group collaboration with multiple screens, cameras, and tables. Figure 1 shows the hardware design of the system. A single t-Room consists of eight modules (called Monoliths) arranged in a nonagon (one side left open to enter) with a shared worktable at the centre which consists of two LCD displays. Each Monolith is made up of a 65 inch LCD panel (resolution of 1920 by 1080) and a high-definition video (HDV) camera mounted on the panel. Each camera aims at a LCD panel standing at the opposite side of the room and captures a participant's whole upper body standing in front of the panel. A polarizing film is placed over each camera lens to capture only the views in front of the display and to eliminate infinite video feedback. The captured image is displayed on the remote t-Room's Monolith placed in the corresponding position (a mechanism similar to [29]). The configuration of video cameras and screens maintains spatial relationships between distant sites, and is designed to support consistent spatial referencing as well as a coherent environment in which to accomplish interaction. Because of this simplicity, however, an object located near the boundary of a Monolith is inevitably captured by more than one camera from various angles. Such a tendency increases as the object gets closer to the centre of the room, resulting in multiple object images being simultaneously displayed on multiple Monoliths.



**Figure 1. Hardware design of t-Room:**  
Top-view (left) and Bird's eye-view (right)

Similar to the Monoliths, a central worktable consisting of two 40 inch LCD panels with two HDV cameras hung from the ceiling captures the activities above the table. Again, polarizing film is placed over each camera lens to eliminate video feedback. For verbal communication, we captured local sound with wireless microphones worn by each participant and played on the speakers in the remote t-Room. In this study the audio was not spatialized.

In addition to the providing views of the co-participants and the documents on the tabletop, t-Room allows data sharing across sites; distant people can share digital contents such as moving images displayed on one of the

Monoliths. For this study a simple graphical user interface was developed on a notebook PC to allow participants in each t-Room to control what was presented at both sites. By storing the contents in each location ahead of time, people could view the same materials almost at the same time. Furthermore, by displaying the remote activities over the digital contents, the system allows people to collaborate over, around and about the digital materials.



T-Room in Kyoto, the two participants in Tokyo are presented on the left, they and one participant in Kyoto, are pointing at the 'shared image'. When pointing over the image the remote participant's conduct is overlaid.

The view from Tokyo at the same time (from a camera above and to the side of the shared screen)

## EXPERIMENT & METHODS

Two identical t-Rooms were installed in the cities of Atsugi (near Tokyo) and Kyoto, which are approximately 150 miles apart, and connected by a gigabit network. We used the network to transmit HDV and audio data. The network delay for video and audio transmission between the two cities was around 0.3-0.4 and 0.2-0.3 seconds, respectively; video and audio were not synchronized.

To explore how people collaborated over and around digital materials using the system, we organized a series of quasi-naturalistic experiments that mimic the work practices in an e-science domain. 5 groups of four English-speaking subjects (13 male and 7 female) were recruited for this study and were paid for their participation. For each experiments two participants were located in Kyoto and two in Tokyo. The experiments were designed to examine:

- How participants coordinate their activities through the technology and the difficulties they faced.
- Whether participants could discuss details of the complex images, particularly how conduct is shaped with regard to moving images.
- How conduct (pointing) is designed for both co-located and remote recipients and whether and how the technology transforms this.

We developed a number of tasks for subjects to undertake in the experiment that would reflect the kind of complex

activities required when supporting complex e-research activities: identifying objects from within complex scenes, locating features in dynamic images and reasoning about what is being viewed. As these needed to be generic and not require particular expertise we designed a task based around the analysis of clips from films of Alfred Hitchcock. The tasks the participants were asked to do ranged from trying to find Hitchcock within a clip, counting people who were looking at a particular feature and mapping out the environment in which a scene takes place. These problems were modeled on the kinds of activities that are undertaken by social scientists when analyzing video materials and included transcribing vocal and visual conduct, mapping out gaze direction and trying to discover the sequences of activities in fragments of data. To encourage discussion over the table, we also prepared some tasks using paper documents. The participants were given a 10 minute introduction to the technology and the task lasted around 90 minutes. There followed a short de-briefing to collect the comments of the participants about the use of the system.

We collected materials from 6 cameras (3 from each t-Room). Our analysis draws on conversation analysis and recent studies of multi-modal interaction. In this respect our concern is with the emergent and sequential character of practical action and the practices in and through which participants collaboratively accomplish particular activities, in this case pointing and reference. The fragments presented here are selected in order to provide a sense of some of the issues and complexities that arise in these brief moments of concerted action. In this paper we focus on the activities of the participants that arose when they were exploring and examining the moving images.

### POINTING FOR MULTIPLE PARTICIPANTS

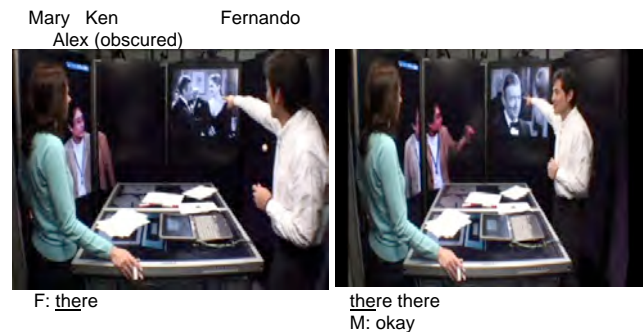
The tasks did seem to provide all four participants with the opportunity to engage in collaboratively finding objects and identifying features in the moving images. When a participant did locate a feature they would typically point to it on the central screen, both to their colleagues at the remote site and to their co-present, co-participant. In the following fragment, the participants are trying to find Hitchcock in the party scene from the film *Notorious*. Alex (A), in Tokyo, has made one suggestion which has been rejected. Now Fernando (F), in Kyoto, notices a figure who looks like Hitchcock and points to him on the screen to his right.

#### Fragment 1 (Clip 8)

F: The- there there  
(0.3)  
M: okay that is him.  
F: is that him↑  
(1.2)  
A: Yeah.

As a bald man appears on the screen Fernando, who is controlling the playback utters 'the- there there' and

moves towards the screen. As he does this he holds his pointing towards the screen, even as the shot changes.



Fernando turns to his co-participants in Tokyo, Alex and Ken (K), and seems to get some acknowledgement from them (both nod, Ken also points to the screen and Alex says 'yeah') and also from Mary (M) who is standing to his right in Kyoto.

Fernando then goes on to discuss whether the man he has pointed to is indeed Hitchcock. Fernando's conduct, a simple pointing to the screen, secures engagement from all three co-participants. He identifies a feature for his colleagues, they seem to recognize what is being pointed at and Fernando monitors that they have located that feature. This provides a foundation for quite a long discussion about whether this is or is not the same man as one they have previously rejected. However, Fernando's conduct is produced in distinct ways for each participant. Once he has moved towards the screen he turns towards Alex, keeping his hand held out and asking 'is that him?', only dropping it when Alex nods and says 'yeah'. As Fernando moves back to where he can operate the playback he turns briefly to Ken, who now appears oriented to the film, smiling at the screen. Although she is in the same (t-) room, Fernando does not orient directly to Mary. Fernando, therefore, distinguishes the recipients of his conduct, securing different qualities of agreement from each.

T-Room, in quite subtle ways, allows participants to shape the 'production format' of their conduct, segmenting the 'audience' into principal and ratified recipient(s) [cf. 12]. This can be accomplished through the articulation of a turn of talk and the production of what seems a simple pointing action. It is also done through the configuration of the body. Fernando moves close to the screen and holds his arm whilst oriented primarily to Alex (but also to Ken). He then holds this pointing, monitoring their response. When this is secured he withdraws his hand. The accomplishment of the point relies on the co-participation of his colleagues. In this case, the recipients respond in distinct ways. Mary seems sensitive to this and from her position on the left turns towards the images of Ken and Alex and then explicitly says while pointing at them 'They're in agreement↑ (0.3) Right'.



Mary turning to screens behind her

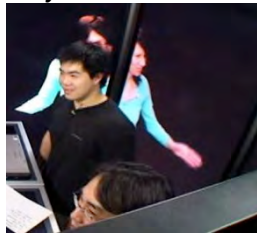


M: They're in agreement↑ Right.

Considering the activity as a simple point that merely identifies a referent for a co-participant, glosses over the collaborative nature of the conduct. Moreover there are other interesting features of its production. First, as Fernando is pointing to a moving image the actual 'referent' of his point disappears almost as his gesture reaches its full extent. Fernando seems to adjust his pointing in the light of the changing image, pulling his finger in whilst still holding his hand out to the place where the feature appeared. This still secures a response from Alex. Second, even given the fast connection speed there are slight delays between the production of the conduct at one site and its appearance on the other, further weakening the ties between the pointer and the object he is pointing at. Third, the environment of action is also subtly transformed. In t-Room there are positions where a remote participant can appear to be looking at you even though he or she is looking at your colleague who is in the same space. This may account for Ken's nodding when Fernando seems to be addressing Alex. More noticeably, when Mary turns behind her to look at Alex and Ken in Tokyo, her own image is not visible to them, as her image (for them) is behind them.

**Kyoto**

Mary turning towards Alex who appears behind her

**Tokyo**

Mary turning to Alex, appearing behind him

T-Room offers high fidelity representations, embodiments, of the participants in an environment where participants have a great deal of choice about how they can position themselves and shape their own conduct. Even so the technology transforms how that conduct appears at the remote site in some subtle (and some not so subtle) ways. Nevertheless, it is curious that throughout the tasks the participants manage to recover an appropriate sense of their colleague's actions. Indeed, they did not report problems when asked afterwards about their impressions of the system or anything they noticed as strange, and during the tasks they made very few explicit requests to clarify another's conduct. In the next sections we will consider how the participants seemed to produce and

make sense of each other's conduct in this multiparty environment.

**RE-PRODUCING A POINT**

As the resources the participants have in t-Room are in some ways symmetrical, it is often the case that participants identify the same feature at the same, or close to the same, time. As in the following fragment where the participants are trying to find Hitchcock in a street scene. Soon after the image fades in, both Andrew in Tokyo and Helen in Kyoto seem to spot Hitchcock in the centre of the image, standing on the pavement by a post box.

**Fragment 2 (clip 56)**

- A: Yeah Right (0.2) there STOP:  
(0.6)  
H: ah him  
F: uh huh  
A: (.) right (.) up (.) No here let me (scoot around) from you (Brian) that guy right there exactly↑  
(0.2)  
A: that's got to be him  
H: okay alright we have finished the clip and could not find anyone else

As Hitchcock appears, Andrew who is standing furthest from the screen points to a man in a black coat. He leans across his colleague Brian, stretching almost as far as he can go. As he says 'stop'. Brian, who has been writing some notes looks to the screen where Andrew is pointing.

**Tokyo**

Andrew Brian Helen Francis  
(Obscured) (out of frame)

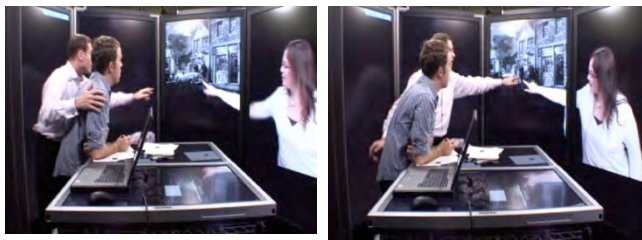


A: there

STOP: can you stop it

Andrew's 'stop' is for Francis in Kyoto who is operating the playback controls and Andrew's conduct seems to get both Francis and Helen to look at the screen and find who he is referring to. Indeed, as Helen who is closer to the screen, turns, her eyes open wide and her mouth opens, as if in surprise at identifying Hitchcock. She goes on to point herself to the man in black, saying 'ah him'.

By holding his point to the screen for nearly three seconds, Andrew gets all of his colleagues, both in Tokyo and Kyoto, to orient to the screen and identify what he is pointing at. Helen also produces a point to the screen. However, there seems to some problem with the precise location to where Helen is pointing. Andrew after suggesting that the location Helen is pointing to is problematic ('right up no'), actually goes on to walk around behind Brian and almost touch the screen when he points at the figure again.



A: right up no, let me (scoot around) from you (Brian) that guy right there exactly↑

As Andrew reproduces his gesture, Helen also points again. They both now move their fingers up and down and around the same figure on the screen. This time they seem to agree that they have found their man. The difficulties, for Andrew, of his earlier pointing seem to be due to the way that his pointing gestures have appeared in the different domains. There are slight differences in how they appear to the pointer in the local domain and the pointee(s) in the remote one.



For Helen (bottom) her gesture in the light of Andrew's which is away from the screen, seems to more precisely identify the object, and yet, for Andrew (top) Helen's finger seems a little lower and is directed towards a car that is moving across the scene; for Andrew, it is more directed towards the driver. Both participants seem to be sensitive to a problem in some way: Andrew goes on to more precisely identify the person he has found, reformulating his own prior pointing and also implicitly addressing problems he sees with Helen's gesture. On the other hand, Helen goes on to reformulate her own prior conduct, also moving closer to the screen and circling around the man on the screen. Although this appears slightly lower to the participants in Tokyo it does serve to unambiguously identify the object.

T-Room by careful calibration and integration of digital images allows participants' conduct to appear to another participant hundreds of miles away apparently over the same object on a common screen. For most purposes, where the object in question can be disambiguated through talk and the general orientation of a look or body movement, this is adequate. But when there are possible alternatives more precision is required. Even for what seems like a simple task, participants do sometimes require quite detailed precision in the production of their visible conduct, and are very sensitive to how this is being oriented to by a colleague. Systems like t-Room, compared to the early media spaces and conventional forms of video-conferencing, offer many more and richer resources for allowing participants to identify objects in a remote domain, and perhaps more importantly, they do seem to support seeing others' conduct with respect to those objects. However, there still are asymmetries in how one's conduct appears to another. Although they may not be aware of the details of these, of how their conduct appears to another, participants are aware when these problems arise and take remedial action, like moving closer and repeating a point to an object, to resolve the ensuing difficulties. In t-Room such common sense solutions help resolve some of the problems of calibration. Although reproducing a point closer to the screen may not necessarily make it clearer, indeed it may exacerbate the problem, it does allow the pointer to re-produce the gesture in the light of that of the remote colleague. The symmetrical resources offered by the system allow participants not only to self repair and to repair the conduct of a colleague but also to monitor how the other orients to that corrective work. The selfsame resources for producing and recognising action are used to repair and resolve problems that seem to arise.

**CO-PRODUCING A POINT**

Most of the activities the participants were set involved quite complex problems, including mapping out what was happening in a clip, transcribing the talk in a scene and even analyzing the gaze direction and visual conduct of the characters in the scene. To solve these a division of labour frequently emerged between the participants. In the following fragment the participants are trying to work out how many people in a restaurant notice that an argument is occurring between the main protagonists. The scene is complex, full of people, and also involves a pan as the principal characters move across the room. Therefore, various objects move in and out of view. To address these complications the participants in the experiment have taken on different interrelated activities. Adam in Kyoto is operating the playback controls and keeping count, Michael furthest from the screen in Tokyo is looking out for candidates which Jenny in Tokyo and John in Kyoto, take turns to point out. Jenny and John are both closer to the screen and can also look out for anyone the others might have missed. They are coming to the end of the

scene, and John has just pointed to a sixteenth woman, when Michael identifies one more person John seems to have missed, 'and the guy behind her', and accompanies these with a point to the screen.

View  
of  
Tokyo  
site  
from  
above



Michael Jenny John Adam  
M: and the guy behind her (partly out of shot)

However, because of the way it is displayed through t-Room, in Kyoto, this conduct is hardly noticeable to John.

View of  
Kyoto  
site  
from  
above  
screen



John Adam Michael  
M: and the guy behind her

For Michael (and Jenny) in Tokyo, Michael's pointing reaches out towards the screen, but for John and Adam not only is Michael's conduct fractured by the border between the Monoliths, it is presented behind them and rendered in two dimensions. John remains looking at the screen. Jenny who has both seen Michael's gesture and can assess John's response, points to the screen.

Tokyo



Michael Jenny John

John then nods, and Adam says 'that guy there, seventeen'. The participants thus not only distribute the work of solving the task between the two sites, but also can draw on the resources available to resolve problems due to the nature of the way images are presented in t-Room. The participants have no explicit resources for seeing how their own conduct is being presented in the remote domain. They can see how the images of their remote colleagues appear and perhaps draw on this to design their own conduct, but the high fidelity and the life-size nature of the images gives few clues to the ways in which these images have been subtly transformed. Rather, the participants seem to draw on the ability to continuously

monitor the responses of a co-participant to their own actions, and if necessary repair or reformulate any problems as they emerge. They also do this in the light of the responses of the other participants. So, in the case above, Jenny not only displays a sensitivity to both John's and Adam's apparent failure to respond, withdrawing her hand when John reorients to the place she is pointing. It was thus not uncommon to see a form of extended collaborative pointing, where a co-present colleague nearest the screen reproduces a point just made by a colleague. The pointing, or more accurately the identification of the object, was co-produced by more than one participant in the light of the conduct of the remote participants many miles away.

### REVEALING THE QUALITIES OF AN OBJECT: ANIMATING FEATURES

By offering the capability to present an image of the entire upper body t-Room allows participants to produce a variety of gestures to animate features of the objects they are describing. These can often accompany quite complex descriptions of what they see on the screen. In the following fragment, participants are trying to map out the room in the famous auction scene from *North by Northwest*. They have to locate where the principal characters are positioned during the action. To help them they have an outline plan of the room on the desks in front of them. Alex, in Tokyo, suggests that Cary Grant is sitting in the centre of the room. In order to find out exactly where this is Alex proposes matching the number of rows of seats visible in the scene with those on the paper plan. Ken, his colleague in Tokyo, and him do so, counting out aloud, with Ken who is closer to the screen animates the counting by pointing to the screen as they do so. Ken's conduct is shaped towards Fernando in Kyoto, who responds with a concern about this suggestion.

#### Fragment 4 (Clip 27)

- F: But he's he's looking very:: like straight , forward no↑  
(0.8)  
F: So he's not here in the in the middle he's probably here no?  
(1.5)  
F: what do you think?  
M: there's an aisle, right↑ on the side here ( )  
(0.2)  
F: ya ther- there's an aisle here (1.7) but he (0.2) he's (2.1) if  
he would be here in the middle his face would be (0.8)  
F: a lot more (.) tilted than (3.0) what do you guys think?

As he say's 'But he's he's' Fernando moves closer to the screen and then gestures over the image whilst suggesting they use where Cary Grant is looking to help locate where the actor is sitting (i.e. in relation to the auctioneer's rostrum which Grant is looking at).

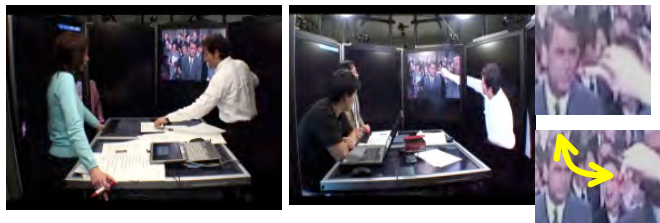
Fernando's gestures animate the gaze direction of the actor, first pointing at the actor and then with an open hand moving it back (whilst saying 'like straight') and



he's looking very like

straight forward no↑

forth (whilst saying 'forward'). As he does this he looks towards Ken and Alex in Tokyo. Although oriented towards Fernando and the screen he is gesturing over, they offer little in the way of response to Fernando's query. Fernando tries again, this time also pointing to the plan on the desk with his left hand and looking directly at the pair in Tokyo. Alex and Ken both look between page and screen but still offer no response. After he tries a third time, explicitly asking what they think, Mary also contributes and gestures to show the location of an aisle both on the screen and on paper. Fernando tries again, this time using his pen. Whilst he points to the centre of the worksheet with his left hand, Fernando says 'if he would be here in the middle his face will be a lot more tilted', he animates the direction of gaze as if the actor was in Ken and Alex's suggested position. Fernando moves his pen in an arc towards the right and then towards Alex and Ken, tilting his pen up and down.

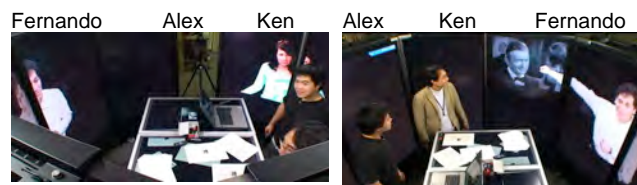


Alex then looks down to the page in front of him and goes on to discuss with Ken, the problem Fernando has raised, that the two ways of determining the actor's position produce inconsistent results. The colleagues in both sites together go onto to review the fragment to see if there is a way of resolving the problem.

In this fragment the four participants are engaged in a very complex form of reasoning. They are attempting to assemble a coherent account from the various images they can jointly see, drawing upon the perspectives of those characters that appear in the images. This kind of argument would be familiar to researchers who analyse naturalistic video recordings of social interaction in data sessions [30]. It also resonates with the practices of researchers working with complex visual data [6, 13]. Researchers when working with complex data juxtapose details of images with other resources such as representations on paper, utilise other material artefacts like pens and pencils to make more apparent particular details of what is visible, and produce complex animated gestures to reveal aspects they consider significant, even if these features may not be directly available from the

image [6]. Through the 'embodiments', t-Room provides participants with a range of ways of animating an image or an argument. However it is not just that the conduct of a co-participant is re-produced in high fidelity or that most of the body appears visible [26], t-Room allows participants ways of relating their conduct to features of the environment, in the light of the orientation of their co-participants. Fernando can monitor the ongoing engagement of his co-participant(s) and reshape his own conduct in the light of this, reformulating his explanation and drawing upon other kinds of resources. It is not just the ability to reproduce an embodiment that is critical but also to provide a coherent relationship between that embodiment, its environment and the contributions of others.

T-Room transforms the environment of action in other ways. As alluded to earlier the configuration of the Monoliths means that participants, even in the same space, can have a different sense of their remote participants. At the same time the remote party may appear to be looking at you, whilst they are actually oriented to your colleague, the so-called 'Mona Lisa' effect.



Tokyo: from behind the screen  
Fernando appears to look at Ken

Tokyo: From in front of screen  
Fernando appears to look at Alex

Moreover, transmission delays often resulted in the images of a remote participant appearing some time (up to ½ a second) after it was produced. At times this means that an action by a remote party, that was produced prior to one's own conduct, actually appears to be responsive to it, with the reverse happening at the other site. So, in effect, parties at both sites in the interaction can consider their own action as having initiated an activity with their colleagues seeming to respond appropriately. Curiously, neither party would comment on this nor did this seem to evoke the parties to try and repair any misunderstanding. T-Room transforms both the spatial and the temporal environment in which interactions are accomplished. In prior studies the timing of conduct in video-mediated interaction [27] and the spatial configuration [11] have seemed to be critical and have led to sophisticated solutions to try and resolve them [e.g. 31]. Faster communication speeds have no doubt overcome some of the problems caused by delays and higher resolution images given greater access to a remote environment, but one might still expect that, particularly for tasks that require the precise location of features at the right time, these transformations would be problematic. Rather, participants seem to make sense of their co-participant's activities and implicitly self correct their own ongoing



conduct. The high fidelity of the system is therefore more essential in providing support for the moment-to-moment management of situated conduct.

## DISCUSSION

The participants in the experiments were given complex tasks to perform. They had to find, identify and describe objects and features that were small, partly obscured and often indistinguishable from others in cluttered and dynamic scenes. T-Room seemed to support them to do this. The technology is not based on the typical model for high definition, virtual 'meeting' spaces. Rather than being seated around a desk, participants can move around. Rather than one person presenting materials through a display that is distinct from the rest of the space, the shared display is integrated into the interactional environment. Rather than having a simple mechanism for pointing and reference, the aim is to be able to provide a richer range of resources: the whole body can be used to refer to and animate the features one is talking about. Such resources would seem critical when participants need to engage in work with images and discuss and reason about the details of common artefacts.

However, even when developing a space that aims to replicate the ways in which visual conduct is produced and recognized in co-present settings, certain, seemingly minor, transformations are introduced when the distributed technology is designed. These seem to introduce boundaries in the spaces for interaction and fragment how an activity is produced from how it is seen at the remote site. In t-Room these differences do seem to be quite minor. There is a very slight delay between the production and receipt of an image, the flat orientation of the Monoliths mean not only that there are boundaries between the screens but slight discontinuities within the space and as the screens can only present in two dimensions, the shape and impact of gestures and other forms of visual conduct are transformed when seen in the remote location. Although each of these would seem subtle transformations, studies of social interaction would suggest that they should be critical and problematic for participants. The timely production of gestures and visual conduct and the fine details of how they are produced for a participant from moment-to-moment have been a recurrent concern in studies of naturalistic conduct [14]. However, in t-Room the participants do not seem to face many difficulties overcoming the effects of these transformations to the interactional space: they rarely explicitly refer to the discontinuities in the timing and shape of another's conduct and did not report noticing these when asked afterwards about the system. Indeed, when certain aspects of the conduct were demonstrated to them afterwards many found it just as curious. The participants seemed to manage to resolve these problems in the course of the production of their activities. In future work we plan to undertake further experiments to explore when the delays and further alterations to the spatial

configurations become problematic in multi-party interaction.

In a range of systems researchers have explored a variety of ways of supporting the ability to refer to objects in a remote environment. This may be through the use of laser pointers, robotic arms or avatars that can gesture in CVEs [17, 22]. Studies of these techniques have revealed how it is problematic for participants to refer to objects, or at least for others to recover what is being pointed at [17, 23]. These technologies seem to fracture and fragment the environment, disrupting the sequential and emergent ways in which pointing and related kinds of visual conduct emerge. Where this seems to be preserved, say when other displays provide resources not only to see the trajectory of the conduct but to monitor how a co-participant is attentive to that trajectory, a more coherent environment of action seems to emerge [24]. T-Room provides such an environment: a co-participant can see the initiation of a gesture, gear their own conduct accordingly and the person accomplishing that gesture can also monitor the co-participants' ongoing assessment to their own conduct. The 'embodiments' provide the 'feedback' necessary to be able to readjust, reformulate and even repair how an action is produced in this distributed environment. Further forms of feedback, like vanity monitors, reproductions of their own images or conduct or mere pointers seem unnecessary, and potentially disruptive. Instead participants draw on the ongoing response of a colleague to reformulate and transform action – even in its course. However, some way of preserving aspects of this referential work, for example by being able to mark up the image being referred to [cf. 7] so that annotations are preserved might serve as a useful additional resource for further discussion and collaboration.

It should be noted that these abilities to resolve referential problems are not just because life sized-embodiments are presented in real-time, but because these embodiments are embedded within the environment of action. It is the embeddedness of the action, not the embodiment of the actor, that is critical. In future work although we intend to develop more 'lightweight' approaches, using fewer (and perhaps smaller) screens for example, to assess the extent to which embeddedness can be supported with less emphasis placed on representing embodiments.

The issues arising from this study seem to resonate with debates concerning the early analysis of interaction in media spaces. The asymmetries in interaction in these early systems did seem to engender transformations in conduct of the participants in these spaces. Participants had problems referring to common objects and talk and visual conduct did seem to be disrupted. Practical matters like assessing the gaze direction of a co-participant and accessing the materials in the remote domain were problematic [16]. And yet it was argued that over time

many of these interactional problems could be resolved and new practices would emerge that enabled participants to interact with each other more easily [3]. Indeed, it did seem that in some cases, ways of looking through the system were transformed and the technology could be made at home in the world. The t-Room experiments cast these early experiences in a new light. The participants did seem to manage the incongruities in the environment, and yet did so through the course of their own actions, almost as a 'problem' emerged they developed ways of resolving it. It seems rather than gradually develop a range of solutions over time, the participants draw on assumptions about their and their colleague's standpoints to resolve the interactional incongruities they faced [cf. 9].

Although it seems like a simple action, pointing is a challenging activity to support when designing collaborative systems. Participants reveal in the course of a momentary action not only their relation to their own environment but also display a sensitivity to how another can see and understand that conduct, as if they themselves were in the other's place. It is perhaps understandable that despite over 20 years of technical development this remains a problem for collaborative system design. Efforts that fragment everyday conduct across distinct domains bring to light what we still need to understand about how people manage and co-ordinate actions in human interaction.

#### ACKNOWLEDGEMENTS

We would like to thank Keiji Hirata and other t-Room members for their support and Mai Okada for help in running the experiment.

#### REFERENCES

- Broughton, M., et al. Being here: designing for distributed hands-on collaboration in blended interaction spaces. in *Proc of OZCHI 2009*. 73-80.
- Buckingham Shum, S., et al., Memetic: An Infrastructure for Meeting Memory, in *Proc of Conf. Design of Cooperative Systems COOP 2006*: Carry-le-Rouet, France. 71-85
- Dourish, P., Adler, A., Bellotti, V., & Henderson, Hgestures., Your Place or Mine? Learning from Long-Term Use of Video Communication. *CSCW Journal*, 1996. 5(1): p. 33-62.
- Dourish, P. and V. Bellotti. Awareness and Coordination in Shared Workspaces. in *Proc of CSCW '92*. Toronto: 107-114.
- Dutton, W.H. and P.W. Jeffreys, eds. *World Wide Research: reshaping the Sciences and Humanities*. 2010, MIT Press:.
- Flor, G.de la., et al., The case of the disappearing ox: seeing through digital images to an analysis of ancient texts, in *Proc of CHI 2010*. Atlanta. 473-482
- Fraser, M., et al., Remote Collaboration over Video Data: Towards Real-Time e-Social Science. *CSCW Journal*, 2006. 15(4): 257-279.
- Fussell, S.R., et al., Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Human Computer Interaction*, 2004. 19: 273-309.
- Garfinkel, H., A Conception of and Experiments with Trust as a Condition of Stable Concerted Actions in *Motivation and Social Interaction*, O.J. Harvey, Ed. 1963, Ronald Press.
- Gaver, W.W., et al. Realizing a video environment: EuroPARC's RAVE system. in *Proc of CHI 92*. Monterey, CA: 27-35.
- Gaver, W.W., G. Smets, and K. Overbeeke. A Virtual Window on media space. in *Proc of CHI '95*. 1995: 257-264.
- Goffman, E., *Forms of Talk*. 1981, Oxford: Blackwell.
- Goodwin, C., Seeing in Depth. *Social Studies of Science*, 1995. 25(2): 237-274.
- Goodwin, C., Pointing as a Situated Practice, in *Pointing: Where Language, Culture and Cognition Meet*, S. Kita, Editor. 2003, Lawrence Erlbaum: Mahwah, NJ. 217-41.
- Harrison, S., ed. *Media Space 20 + Years of Mediated Life* 2009, Springer-Verlag: London.
- Heath, C.C. and P. Luff. Disembodied Conduct: Communication through Video in a Multi-Media Office Environment. in *Proc of CHI '91*. New Orleans: 99-103.
- Hindmarsh, J., et al. Fragmented Interaction: Establishing mutual orientation in virtual environments. in *Proc CSCW'98*. Seattle, WA: 217-226.
- Ishii, H. TeamWorkStation: Towards a Seamless Shared Workspace. in *Proc of CSCW '90*. Los Angeles: 13-26.
- Ishii, H., M. Kobayashi, and J. Grudin. Integration of Inter-Personal Space and Shared Workspace: Clearboard Design and Experiments. in *Proc of CSCW 92*. Toronto: 33-42.
- Jirotko, M., et al., Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. *CSCW Journal*, 2005. 14 (4) 369-398.
- Kirk, D., T. Rodden, and A. Crabtree, Ways of the hand, in *Proc of ECSCW 2005*: Paris 1-21.
- Kuzuoka, H., et al. GestureMan: A Mobile Robot that Embodies a Remote Instructor's Actions. in *Proc of CSCW 2000*. Philadelphia PA: 155-162.
- Luff, P., et al., Fractured ecologies: creating environments for collaboration. *HCI Journal*, 2003. 18(1-2): 51-84.
- Luff, P., et al. Handling Documents and Discriminating Objects in Hybrid Spaces in *Proc of CHI 2006*. Montreal 561–570.
- Mantei, M., et al. Experiences in the Use of a Media Space. in *Proc of CHI '91*. New Orleans, Louisiana: 203-8.
- Morikawa, O. and T. Maesako. HyperMirror: Toward pleasant-to-use video mediated communication system. in *Proc of CSCW'98*. Seattle, WA: 149-159.
- Ruhleder, K. and B. Jordan, Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. *CSCW Journal*, 2001. 10(1): p. 113-138.
- Tallis, R., *Michelangelo's Finger: An Exploration of Everyday Transcendence* 2010, London: Atlantic Books.
- Tang, J.C. and S.L. Minneman, VideoDraw: A Video Interface for Collaborative Drawing. *ACM Transactions on Information Systems*, 1991. 9,(2): 170-184.
- Tutt, D., et al. The Distributed Work of Local Action: Interaction amongst virtually collocated research teams. in *Proc of ECSCW 2007*: 199-218.
- Vertegaal, R., et al. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. in *Proc of CHI 2003*. Ft. Lauderdale, Florida, USA: 521-528.3.