

# Embedded Interaction: The Accomplishment of Actions in Everyday and Video-Mediated Environments

PAUL LUFF, King's College London

MARINA JIROTKA, University of Oxford

NAOMI YAMASHITA, NTT Communication Science Laboratories

HIDEAKI KUZUOKA, University of Tsukuba

CHRISTIAN HEATH, King's College London

GRACE EDEN, University of Oxford

A concern with “embodied action” has informed both the analysis of everyday action through technologies and also suggested ways of designing innovative systems. In this article, we consider how these two programs, the analysis of everyday embodied interaction on the one hand, and the analysis of technically-mediated embodied interaction on the other, are interlinked. We draw on studies of everyday interaction to reveal how embodied conduct is embedded in the environment. We then consider a collaborative technology that attempts to provide a coherent way of presenting life-sized embodiments of participants alongside particular features of the environment. These analyses suggest that conceptions of embodied action should take account of the interactional accomplishment of activities and how these are embedded in the material environment.

Categories and Subject Descriptors: H.5.3 [Group and Organization Interfaces]: Synchronous interaction

General Terms: Human Factors, Design

Additional Key Words and Phrases: Media spaces, video-mediated interaction, reference, embedded interaction

## ACM Reference Format:

Luff, P., Jirotko, M., Yamashita, N., Kuzuoka, H., Heath, C., and Eden, G. 2013. Embedded interaction: The accomplishment of actions in everyday and video-mediated environments. *ACM Trans. Comput.-Hum. Interact.* 20, 1, Article 6 (March 2013), 22 pages.  
DOI: <http://dx.doi.org/10.1145/2442106.2442112>

## 1. INTRODUCTION

There is a long-standing interest in developing technologies to support real-time collaborative work between people in different physical locations. Initially, through video telephony and later through video conferencing systems, the development of novel technologies to mediate work and communication has resonated with the increasingly globalized nature of work and the provision of services and seems to support contemporary organizations and their inter- and intra-institutional arrangements. Over the past

---

This research was partly supported by Embedding e-Science Applications-Designing and Managing for Usability project grant no. EP/D049733/1.

Authors' addresses: P. Luff (corresponding author), Work Interaction and Technology Research Centre, King's College London, UK; email: [paul.luff@kcl.ac.uk](mailto:paul.luff@kcl.ac.uk); M. Jirotko, Department of Computer Science, University of Oxford UK; N. Yamashita, NTT Communication Science Laboratories, Kyoto, Japan; H. Kuzuoka, Division of Intelligent Interaction Technologies, University of Tsukuba, Japan; C. Heath, Work Interaction and Technology Research Centre, King's College London, UK; G. Eden, Department of Computer Science, University of Oxford, UK.

Permission to make digital or hard copies of part or all 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 show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2013 ACM 1073-0516/2013/03-ART6 \$15.00

DOI: <http://dx.doi.org/10.1145/2442106.2442112>

couple of decades we have witnessed the emergence of a range of solutions, sometimes characterized under the auspice of “media spaces”, designed to enhance interpersonal communication between distributed participants [Harrison 2009]. Notwithstanding these initiatives, it is widely recognized that the application of these solutions has been relatively limited and systems have failed to provide personnel with access to each other and the respective resources on which many of their workplace activities rely. As we have remarked elsewhere, a “face-to-face model” of interpersonal communication has underpinned many of these developments, a model exemplified in the head-and-shoulders frame used in many video telephony and video conferencing facilities [Heath and Luff 1992b]. Recently there has been an interest in enriching the access to participants have to each other in media space, where systems have been designed to enable people to see each other’s bodily comportment and determine their spatial configuration within the distributed environment. In other words, there has been a growing commitment to supporting embodied action, multimodal communication, between remote participants.

Alongside these initiatives, there has been a growing recognition of the importance of providing participants not only with access to each other, but also to the material and digital resources on which their work often relies. Increasingly, both commercial and experimental systems enable participants to see and access each other’s documents, both digital and material, and in some cases provide ways to annotate those documents during the course of their mutual activity. The contributions of these initiatives, however, have to be seen in the light of what we know of work and collaboration in conventional organizational environments and the rich variety of resources on which people rely in working together. The growing corpus of research that has come to be known as workplace studies has not only served to demonstrate how all sorts of objects and artifacts, tools and technologies, feature in the collaborative production of organizational activities, be they electronic documents, slips of paper, keyboards, telephones, pencils and pens, and the like, but revealed how these resources are contingently deployed and used within the developing course of the participants’ interaction. They demonstrate the reflexive, mutually interdependent relationship between action and the occasioned features of the local environment; interdependencies that inform both the production and intelligibility of action. In other words, it is not simply that to produce a particular action one may rely only upon, for example, looking at a particular document or using a particular technology, but also the ability of others, with whom one is working, to recognize just what one is doing relies upon their ability to determine the reflexive relationship between the object and action. In other words, the sense and production of action is inextricably embedded within occasioned features of the local ecology; occasioned by virtue of, and within, the emerging activity and interaction. The expression “embodied” is sometimes used to capture a sense of both the bodily and ecological character of social action [Dourish 2001; Robertson 1997], but like the term multimodal, it can inadvertently draw attention away from the ways in which the production and intelligibility of action is entailed and dependent upon occasioned features of the immediate environment in which it occurs. In this article, we seek to suggest that an important limitation in the development of systems to support distributed collaborative work derives from their impoverished conception of the embodied, and in particular, the embedded character of practical action.

If there is one collaborative activity that exemplifies the embedded character of practical action then it is reference, and in particular, pointing. As workplace studies powerfully demonstrate, much collaborative activity relies upon the participant’s ability to unambiguously refer to and point at features of objects within her immediate environment. With regard to reference and pointing systems designed to support distributed collaborative work this seems particularly problematic, not simply by virtue of

the limited access they provide to the environment of the other, for example, their office and the resources the other has at hand, but with respect to the ways in which they delimit access to the person (and the person's action) with regard to that environment. There have been a number of attempts to address and resolve this problem, for example, by developing systems that enable remote participants to manipulate cameras in the coparticipants' environment [Gaver et al. 1993], Collaborative Virtual Environments (CVEs) in which participants have access to avatars and a shared world [Fraser et al. 2000], or even providing participants with lasers or robots to point towards objects [Kuzuoka et al. 1994, 2000; Yamazaki et al. 1999]. These and a range of other solutions, however, remain problematic; they provide limited access to the respective environments and in some cases exacerbate the fractured or fragmented character of the different ecologies. In other words, the more one enhances mutual access, the more difficult it becomes for participants to embed the actions of the other within the relevant occasioned features of the environment.

This article forms part of a program of research, in which we have attempted to build systems to support the embodied and embedded character of practical action and enable remote participants to unambiguously refer to occasioned objects within a mutually accessible environment. Drawing on a study of activities that arise within particular work settings, we discuss one system, namely t-Room, that has been designed to establish a shared environment to enable participants not only to see a shared world, but to see each other and each others' actions with regard to that world. In this way the article suggests that by prioritizing the embodied and embedded character of practical action we pose severe challenges for those with an interest in developing systems to support collaborative work. In turn, these systems, and prototype solutions, that do seek to support embedded action raise some interesting questions for our analysis of the ecological foundations of everyday practical action.

## **2. EMBODIED ACTION: WORKPLACE STUDIES AND THE DESIGN OF INNOVATIVE TECHNOLOGIES**

Since the late 1980's a wide range of analytic and theoretical perspectives have been brought to bear on the study of embodied action in fields of HCI and CSCW, including perspectives emerging from philosophy, psychology, and the social sciences. From these an impressive empirical corpus of studies has emerged that not only considers verbal and textual communication with and around technology, but also the visual conduct of participants and their use of everyday physical objects. So, researchers have drawn upon phenomenology [Dourish 2001; Robertson 1997], distributed cognition [Hutchins 1995], and course of action analysis [Filippi and Theureau 1993] to consider the ways in which bodily orientation is a resource for supporting communication and collaboration, how apparently mundane tools support complex computational practices, and how the layout and configuration of an everyday environment can facilitate or undermine the activities that are engaged within it. A large number of these studies have been informed by developments in ethnomethodology and conversation analysis [Garfinkel 1967; Sacks 1992]. These include studies of such diverse settings as surgical operations [Koschmann and LeBaron 2003; Mondada 2001; Sanchez Svensson et al. 2007], transportation control centers [Goodwin and Goodwin 1996; Harper and Hughes 1993; Heath and Luff 1992a], financial trading rooms [Heath et al. 1994], medical consultations [Greatbatch et al. 1995], design practices [Suchman 2000], and more general office work [Anderson et al. 1989]. Such studies are undertaken typically using fieldwork, often supplemented by analysis of video recordings to reveal how activities are produced with respect to the contingencies and circumstances of the participants within organizational settings, and examine how the technologies available in these domains are utilized, whether these are simple tools or documents, conventional systems, or

more advanced computer technologies. The studies reveal detailed ways in which the use of technology is collaborative and produced in interaction.

So, for example, analyses of how participants' visual orientation towards a display, which may or may not be accompanied by talk, have suggested how participants in complex settings, like control rooms, can not only help maintain "awareness" of the happenings in and around a domain for themselves, but also provide others with the resources to identify for themselves issues that may be of consequence to their work [Goodwin and Goodwin 1996; Heath and Luff 1992a]. Or, analyses of the details of material conduct reveal how documents can be easily read, or written, in ways that are sensitive to an ongoing interaction, whether the document is a medical record within a doctor-patient consultation or a plan being discussed by a group of architects [Luff and Heath 1998; Luff et al. 1992]. Participants can shape their conduct, their talk, body orientation, and gaze with respect to shifts and transitions in the conduct of their colleagues [Luff et al. 2004]. Or, consider a seemingly simple activity such as when someone refers to an object or a feature of an object by pointing to it. This concerning, what is spoken and what is visible is coordinated with the activities of a colleague and also tied to features of the local environment [Goodwin 2003; Hindmarsh and Heath 2000]. Workplace studies that draw upon conversation analysis and ethnomethodology are distinctive not just because they emphasize the collaborative nature of embodied action but also because they reveal how embodied interaction is accomplished through an interweaving of talk, visual conduct, and features of the material environment.

The nature of this conduct has suggested ways in which novel technologies could be configured and enhanced to support collaborative activities when the participants are remote. In early deployments of media spaces, audio-visual environments to support communication, it was noted that conduct seem "disembodied" [Heath and Luff 1991]: gestures did not seem to have the performative impact they had when participants were copresent. Moreover, the configuration of the technology, typically only providing head-and-shoulders views of a colleague, limits both the access to the remote environment and the actions that can be performed in relation to objects in the remote domain. A range of developments, not only in media spaces but in other collaborative technologies, have sought to address these shortcomings. Thus, in collaborative virtual environments not only are representations of coparticipants, embodiments, or avatars displayed, but the actions that these perform are made visible so that colleagues are able to see an individual's actions, like a gesture, in relation to objects and features in the virtual environment. Some CVEs include the ability for avatars to "point" as well as provide views that enable all relevant participants to see a "gesture" together with what it is referring to [Fraser et al. 2000]. In video-mediated systems, "head-and-shoulders" views of coparticipants have been supplemented with images of documents or other aspects of the environment. Designers have proposed ways for a local participant to reach into the remote domain and refer to objects and features within it [Fussell et al. 2004; Gaver et al. 1993, 1995; Ishii 1990; Kuzuoka et al. 1999; Tang and Minneman 1991]. Yet even with these capabilities, enhanced media spaces can seem to be restrictive. To provide greater access, remote cameras and "pointing systems" have been developed that can move around a remote domain and through lasers, roving devices, and robots with electro-mechanical arms, objects can be pointed to, identified, and discussed by participants who are working at a distance [Kuzuoka et al. 1994, 2000; Yamazaki et al. 1999]. Each of these capabilities in some way seeks to enhance the presentation or the capabilities of the embodiment. However, despite these enhancements, referring to objects at a distance has still proven problematic. Experiments with these prototype technologies reveal difficulties faced by the participants when they coordinate their conduct within these environments, in how they identify objects or features of it in a remote environment, and how they arrive (or fail to arrive) at a common orientation

to an object. Remote participants may have difficulties making sense of what seems to be a simple activity such as when a coparticipant points to an object in a mediated environment. The activity becomes fragmented in some way and participants have difficulties tying their colleagues' actions (or the presentation of these) to objects in their local environment [Gaver et al. 1993; Hindmarsh et al. 2000; Luff et al. 2003]. Problems of technologically-mediated referential conduct do not seem to be resolved by providing higher-fidelity embodiments or a greater range of "embodied features". Indeed, curiously, despite providing fragmented views or projections of another, some experimental systems seem more successful in supporting the ways participants refer to objects and identify features of them for another. So, in systems like *Agora*, participants seem to have fewer problems referring to objects and making sense of a colleague's conduct, even though only the images of the hands of a remote colleague are projected onto the desk in front of them, and these images seem distinct from the other displays provided by the system [Luff et al. 2006]. This may suggest that we need to draw on a more refined characterization of how participants in everyday settings manage to accomplish referential activities and how their talk, visual conduct, and features of the environment are related.

### **3. A NATURALISTIC STUDY OF EMBEDDED INTERACTION: ANALYZING IMAGES IN COLLABORATION**

While seeming quite distinctive, the setting we consider in this article can be seen to have many features typical of other workplace domains where participants in the course of their work frequently refer to objects and features of objects. The study we draw upon is of classicists who, as part of their everyday, scholarly work, analyze Roman texts. Originally the study was undertaken within the program of e-research in the United Kingdom that seeks to support scientists, social scientists, and scholars working in the humanities. One objective of this program is to build Virtual Research Environments (VREs) through which researchers who typically work apart can collaborate together [de la Flor et al. 2010a]. These environments could be based on conventional workstations or they could make use of richer forms of video-based communication such as those provided by the *AccessGrid* [Dutton and Jeffreys 2010]. With regard to scholars in the humanities, like the classicists considered here, there seems to be the potential for supporting their analyses of texts through the use of new technologies. Currently, sophisticated computational techniques such as image processing are utilized to enhance images and thus assist them in their work. However, it was not clear how to make the results of these techniques available to the scholars, what kind of computational support could be offered alongside the images (e.g., for magnification or annotation), or how collaboration between scholars might be facilitated, either within a team or between researchers with similar interests who are located in various institutions around the world. Our study was undertaken initially to identify the requirements for a VRE for classicists. As well as observing scholars at work and interviewing them, we video-recorded sessions where two or three classicists met to discuss the interpretation of particularly complex texts.

In the following example of a collaborative analysis session, three classicists are working together on a particular problematic text, written on an ancient wooden Roman Tablet (called the "Tolsum Tablet"). The classicists bring to bear different kinds of expertise for their interpretation: Axel specializes in the study of ancient literature and its meaning, Rupert specializes in the study of ancient handwriting, and James specializes in the study of ancient inscriptions. In this fragment they are trying to analyze what is written on the wooden tablet and are trying to determine when it might have been written. One of a set of images of the tablet, produced through novel image processing techniques, is projected onto a screen. These techniques reveal what

the tablet might look like under different lighting conditions and if features, like wood grain, were removed. By providing different views of the tablet, the writing on it appears quite differently to that found in earlier photographs and suggests new readings of the text. The scholars have been analyzing this text for a while and Axel has become concerned that there do not appear to be any “A”s in the text. Axel suggests a form of marks that might be an “A” and seems to find one in the image.

*Fragment 1 (simplified)*

A: There’s got to be some “A”s in the text, (so) where are they?

R: Yes that’s true.

A: In fact, if you look at, if you look at if you look up here now you can see a very similar one.

R: (right)

A: One there.

R: Yes, yes.

A: I think it is actually there in (.) that form.

R: Yes, yes, so that’s the GARGILIUS?

As Axel starts to say “In fact, if you look at”, he moves toward the screen, followed by Rupert. As he moves towards the image his right arm starts to reach out, his index finger pointing towards the top right of the screen (1.1).

*Fragment 1*

(1.1) Rupert

Axel

(1.2)

(1.3)



A: if you look up here now you can see

a very similar (one)

(1.5) there

R: right

Yes, yes

Axel moves closer to the screen (and towards Rupert) to a place where he is now pointing almost vertically to a location at the top of the screen (1.2). As Axel says “a very similar (one),” holding his arm outstretched, Rupert slightly readjusts his orientation. Following this, Axel traces out with his finger a series of strokes over the image. As he completes the final stroke Rupert utters “right.” Axel then repeats the shape, tracing again with his finger a little more quickly. Rupert then says “yes, yes,” and the scholars go on to discuss both the nature of the mark (it happens to be a form usually seen in later texts) and a reading of the word that contains it: “Gargillius”.

The marks the classicists are viewing are hard to see in the image. Despite the image processing it is still very difficult for experts to be able to distinguish marks from the background, let alone clearly identify letters or the strokes from which they are constituted. The scholars then not only have to identify various marks but also try to get colleagues to see what they have noticed. So, it is not uncommon for them to stand close to the projected images to discuss their interpretations. However, as they locate a feature within a scene, the scholars need to assess how their colleagues are orienting to their own conduct. As Axel produces his gesture with his right hand and while looking towards the feature, his orientation allows him to monitor Rupert’s

own movement towards the screen and Rupert’s orientation to the image. As Rupert nears, Axel extends his arm, locating more precisely the place where part of the mark appears. While holding his arm outstretched, Axel assesses Rupert’s orientation and reorientation to where he is pointing. Only once Rupert seems to have located part of the feature does Axel begin to trace around the whole form, first securing Rupert’s alignment and then his agreement to his proposed interpretation of it.

The production of what seems to be a very simple referential activity, in this case Axel pointing to a feature on the screen, is tied to the emerging talk, so that Axel’s finger arrives over the feature just as he refers to the object: “a very similar one.” This activity is also coordinated with Rupert’s own movement and reorientation: Axel produces his conduct not only so that he can develop quite a complex gesture over the image, but also so that he can assess Rupert’s participation and engagement, and reshape and reconfigure his own conduct if necessary. The conduct is a collaborative accomplishment that emerges through the contributions of both participants. It is a sequential accomplishment, even if it is produced with and in relation to a single turn of talk. Both participants contribute to its production in the course of its emergence. In a turn of talk in fragment 1—“if you look at if you look up here now you can see a very similar one”—Axel not only identifies the feature for Rupert tracing over the shape of the image, but elicits an alignment from Rupert and in doing so, shows how the “A” might have been written by the scribe in the 1st Century AD.

The scholars frequently animate the images in this way. In the following fragment Rupert, while identifying what might be an “E,” suggests how the original scribe might have produced the letter.

*Fragment 2*

R: =( ) he stabbed his (.)  
 stilus in

>And then gone down<  
 ↓bomp

And then he's done it again

A: Mm hmm

R: =Or? (.) (must be) h:he must've have gone  
 s:something like that

A: so it's like, 'E' something?



As Rupert says “and then gone down,” he quickly moves his hand down, pulling his hand away sharply, uttering the word “bomp.” Through a simple gesture and vocal expression Rupert reenacts how the stilus’ pressure could have made the mark and the angle at which it may have been held. Their interpretations are not just a matter of understanding the visual elements of an image, but reveal the material qualities of the physical object and how it was manipulated. The classicists display these through their own conduct, through their talk and visual conduct, this conduct closely tied to

features of the environment. Their referential activity is embodied but relies on being embedded within the environment and for colleagues to see it as such.

Although the activities of the classicists may seem to be very distinctive, there are numerous settings where participants analyze complex materials collaboratively, not only in the humanities, but also in social sciences (in data sessions for example [Tutt et al. 2007]) and in scientific domains, when complex images, scans, x-rays, and the like are analyzed by colleagues working together [Jirotka et al. 2005]. In such domains it is not only important that colleagues identify what they are referring to, the features or objects within images for their colleagues, but also that they see these features in particular ways. When interpreting the detail of a scene or image, participants animate their conduct in different ways: through their talk, bodily conduct, and gaze direction. As they do this they also need to assess the ongoing orientation of their colleagues to what they are animating and showing. It is not only that their embodied action is critical for this accomplishment but how it is embedded in the environment within an emerging course of action.

#### 4. TECHNOLOGIES TO SUPPORT DISTRIBUTED EMBEDDED ACTION

Perhaps the most obvious way of developing technologies to support embodied conduct is through video connections of different kinds, as in media spaces (e.g., Gaver et al. [1992]). However, conventional video conferencing systems and media spaces offer limited access to remote resources like documents. Even if additional document cameras are provided, the images from these are often presented in a distinct area so that gestures made on and around an object by a participant can appear fractured from the conduct that is visible on other screens [Heath and Luff 1991]. Hence, a number of researchers have tried to develop “embodied” enhancements to media spaces, most notably by projecting some features of a participant’s conduct, typically details of the hands, into the remote domain [Fussell et al. 2004; Kirk et al. 2005; Kuzuoka et al. 1999; Luff et al. 2006]. Recently, high-definition and high-fidelity video conferencing systems like HP’s Halo, Cisco’s Telepresence, and Polycom’s Open Telepresence Experience have been introduced aiming to provide more coherent environments for interaction. Utilizing high-bandwidth connections these systems can present real-time, life-sized images of coparticipants with little delay, even when participants are many miles apart. Perhaps more consequential has been the careful design of these and a number of prototype systems, like BISI, which have been termed “blended spaces” [O’Hara et al. 2011]. The designers of such systems have paid careful attention to the aesthetics of the environment, removing as far as is possible visual elements that may be distracting and endeavoring to present a space, similar to a boardroom, that is continuous between two sites. They have also been concerned with how participants appear so that, for example, shifts in gaze direction are consistent in the remote and collocated spaces. This is not only facilitated by encouraging people to be seated along one side of a desk but also by the placement of features, like the legs of a desk, to be positioned so that the location of the participants is relatively fixed and related to the orientation of cameras and displays [Gorzynski et al. 2009; Paay et al. 2011]. Such careful design tries to ensure that not only a turn towards a remote participant has a similar appearance in the local domain as it does in the remote one, but that there are few places where there are blind spots or where a feature in the remote environment appears in more than one space at the same time on the video displays [Paay 2011]. However, this consistency seems difficult to maintain when participants need to refer to objects across the two domains. In the boardroom configuration it is hard to place a “shared display” where documents can be presented. Separating one display in an “information strata” distinct from (either above or below) the “social strata” where the images of participants appear, fragments the conduct between the two [Gorzynski et al.



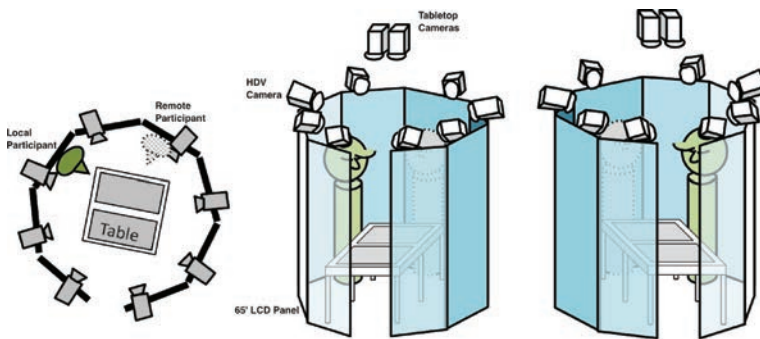


Fig. 1. A diagram of t-Room showing the layout of the monoliths from above (left); and positions of cameras, screens, and presentations of participants in Tokyo (middle); and in Kyoto (right).

2009]. It is harder for the conduct and orientation of the local and remote participants to be consistent, since a gesture towards an object may not appear in the appropriate orientation for all participants. Spaces like BISi try and address this issue by providing multiple “shared displays,” that make the orientation more apparent and also by including tabletop systems to support more focused work [Broughton et al. 2009; Paay et al. 2011]. However, by offering additional document cameras and screens where images from these cameras can be displayed, these spaces are disjoint, so that even a simple action like looking towards or pointing to an object in a remote domain can be fragmented across and appear differently in the various domains provided [O’Hara et al. 2011; Paay et al. 2011]. As O’Hara and colleagues suggest, we need to understand whether these blended spaces resolve problems found in earlier media spaces, particularly by considering the interaction participants have in and through them [O’Hara et al. 2011].

In this light, it may be worth considering the interaction of participants in a blended space that seeks to support ways of integrating embodied actions within the distributed environment. Rather than segmenting the space into distinct “strata” for participants and objects, by mixing images between the domains it may be possible to integrate the images of the coparticipants with the images of the objects. This approach had been adopted by the designers of a prototype high-fidelity system developed by NTT Japan called t-Room [Hirata et al. 2008]. T-Room uses high-definition video, large screens, and image calibration and video-mixing techniques to provide more symmetric resources for remote collaboration (see Figure 1).

A single t-Room consists of eight modules (called monoliths). Each monolith is made up of a large 65-inch LCD screen with a High-Definition Video (HDV) camera mounted on it. The size of these screens and the image calibration allow life-sized images of the remote participants to be displayed in the local t-Room. By configuring the video cameras and screens and calibrating the images, the designers sought to maintain spatial relationships between distant sites, so that activities, such as when participants point to features, are consistent. Polarizing film is placed over each camera lens to eliminate video feedback and also to capture only the views in front of the display. A central worktable in the t-Room consisting of two 40-inch LCD panels with two HDV cameras hung from the ceiling captures the activities above the table, for example, on paper documents placed on its surface.

In addition to the monoliths displaying life-sized presentations of the coparticipants, t-Room allows data sharing across sites; distant people can refer to documents, slides, or moving images displayed on one of the monoliths. Rather than gesturing to a separate display in the remote domain, the participant’s conduct appears so that it overlays the



Fig. 2. T-Room in two sites: Tokyo (left) and Kyoto (right). This image is taken from the experiments. Andrew in Tokyo (on the left) points to a man on the screen which can be seen in Kyoto (on the right). At the same time Helen in Kyoto points to the same man and is visible in Tokyo. It may be noticed that the angles of cameras do result in subtle transformations in the ways the conduct is presented (for example, note the angle at which Andrew's arm appears in Kyoto).

images, he/she is referring to, as if the remote participant is positioned in front of the item (see Figure 2).

Two identical t-Rooms were installed in the cities of Atsugi (near Tokyo) and Kyoto, which are approximately 125 miles apart, and connected by a gigabit network so that High-Definition Video (HDV) and audio data could be transmitted. 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. The speed of the connection and the size of the displays provide ways of presenting life-size images of remote coparticipants in real time. It therefore seems to be a technology that could support the kind of conduct required by researchers, scholars, and other professionals who need to consider and discuss details of documents, diagrams, or moving images. However, although seeming to offer a rich collaborative environment, it is apparent that the technology does transform conduct in subtle and curious ways. There is a slight delay in the transmission of the images from each monolith that is not coordinated with the sound, or with the images from the other monoliths. This means that the production of a gesture (and the accompanying talk) may appear differently to a local colleague than to one at the other site. Moreover, as conduct is displayed on flat screens it only can appear in two dimensions to a remote participant. When discussing details of complex materials these transformations might be consequential. As underpinning the original initiative to develop VREs is an aim to improve collaboration between researchers from different specialties, between distinct disciplines, and across national boundaries, it did seem worth assessing the applicability of the technology to support such activities. Even though it is a large, complex, and expensive system, given its ability to capture and present a rich form of “copresence,” t-Room would seem to offer appropriate capabilities for a VRE. Assessing it might reveal critical barriers for the deployment of, and the key capabilities required in, simpler configurations.

## 5. ASSESSING DISTRIBUTED EMBEDDED ACTION

There are many practical problems of assessing distributed technologies like t-Room. For example, as the sites are many miles apart there are simple problems finding appropriate participants for an assessment which needs to be undertaken in two different locations at the same time. It is unlikely to find a reasonable number of participants who, while having different forms of expertise, have a common set of skills to analyze complex materials. Such materials would have to be relevant to the participants to investigate and explore while away from their worksite. Any tasks the participants would undertake would need to resonate with the activities found in workplace settings, but not necessarily reproduce them. What seemed critical to investigate was the extent to which the technology facilitated the coordination of referential activities through the

technology. The tasks needed to afford opportunities for participants, through their talk and visual conduct, to refer to objects and features in both their own and their colleagues' environments. As in many other cases with prototype technologies, it is often infeasible to deploy these in situ, or assess them in the working environment, with an appropriate set of skilled participants, or even with the types of materials, data, and resources participants use in their everyday work. Therefore, we undertook a quasi-naturalistic experiment where participants engage in an open-ended task or set of activities using a prototype technology [ Benford et al. 1999; Gaver et al. 1993; Hindmarsh et al. 2000; Luff et al. 2003; Yamazaki et al. 2008]. In the case of t-Room, as we were concerned with how the technology might transform collaborative interpretive practices, we required participants to undertake some form of analytic work.

Given it was infeasible to involve a number of participants with a very specific form of expertise, the design of the quasi-naturalistic experiment drew from the study of classicists [de la Flor et al. 2010a, 2010b], and similar kinds of analytic work by other researchers in the humanities [Eden and Jirotko 2012] and by social scientists [Tutt and Hindmarsh 2011]. Although it was recognized that the participants could not draw on very specific knowledge of a domain, the design of the task required the participants would need to not only identify objects from within complex scenes and locate features in images, but also engage in some kind of reasoning about what is being viewed. Through the tasks the participants would need to be able, through their own conduct and the presentations of their conduct in the remote domain, to discuss their interpretations of an image, display their understandings, and contribute to their colleagues' analysis of the materials. More specifically, this kind of analytic work involves the participants in:

- (1) analyzing complex materials together, rather than one individual presenting a previously developed analysis to all the other coparticipants;
- (2) engaging in forms of referential activity other than simple pointing (e.g., counting, animating details, comparing features within and across images);
- (3) developing interpretations that rely on matters that are not directly visible in the images;
- (4) involving colleagues in different forms of participation, including focused discussion of details, engaging in parallel but related activities and undertaking distinct tasks; and
- (5) juxtaposing details of the images with physical objects like paper documents.

Considering how classicists and other researchers engage with colleagues and the materials they have available, it should be possible for the participation in the activity to shift from moment-to-moment.

The materials chosen for the experiment were clips from the films of Alfred Hitchcock. We checked all participants had knowledge of this director's work and designed a set of tasks where participants not only had to locate features but reason about them. Hence, the participants were asked to accomplish a range of activities from trying to find Hitchcock within a clip, distinguishing people who were involved in an activity in some way, discovering the order in which a series of actions happens, attending to the details of sequences of activities, and mapping out the features in which a scene takes place. The activities mirrored the collaborative research meetings we observed. They also required participants to attend and discuss features of complex materials, make notes of these, and consider a range of representations of these materials. Because of familiarity with the work of the director, participants often drew on their knowledge of the films in question in their discussion. The complexity of most the scenes also required participants to pay close attention to the materials being presented and all contributed to the discussions, not just through talk, but also through their visual conduct. The



Fig. 3. Participants reconfiguring their positions in the t-Room experiments, for example, pointing towards the screen from a distance (left); moving close to the screen to identify a detail (center) or gathering around a document on the tabletop (right).

tasks which had been developed required them to refer to quite fine details and hence, engage in a collaborative referential activities.

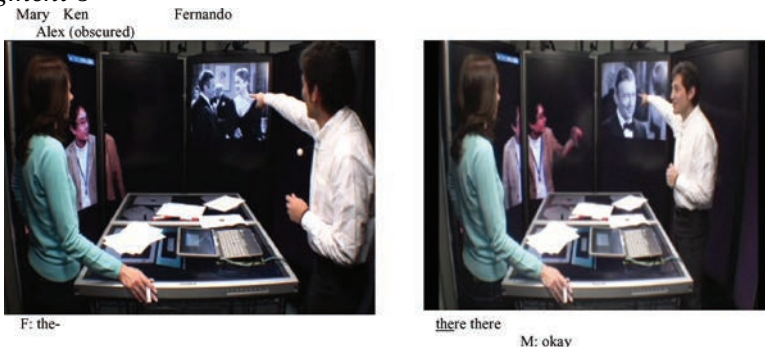
Following pilot experiments we undertook a series of five quasi-naturalistic experiments, each involving groups of four English-speaking participants. The participants were given a 10-minute introduction to the technology and the task, and each experiment was followed by a short debriefing to collect the comments of the participants about the use of the system. The tasks the participants were given lasted around 90 minutes. We collected materials from six cameras (three from each t-Room) and drew on the materials collected to analyze the participants' collaboration through the system [Luff et al. 2011].

## 6. THE ACCOMPLISHMENT OF DISTRIBUTED EMBEDDED ACTION

In each of the experiments the tasks seemed to be engaging and provided opportunities for all four participants to find objects collaboratively and identify features of those objects in the moving images. The tasks also seemed to encourage the individuals to develop flexible forms of participation in the activities; they could move around the space and reposition themselves with respect to the conduct of their colleagues, to the details on the screen, and to features of documents on the tabletop (see Figure 3).

In the course of their activities the participants frequently referred to objects on the screens, not only through their talk but also through their visual conduct, frequently pointing to the screens and gesturing over them. 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. Fernando (F), in Kyoto, has noticed a blurred bald figure in the background who looks like Hitchcock and asks to review the clip to see if “he is going to be there again.” As the bald man appears Fernando utters “the- there there” and moves towards the screen holding his right arm towards the screen.

### Fragment 3



Fernando keeps his arm outstretched, even as the shot changes. Fernando turns to his coparticipants in Tokyo, Alex and Ken (K), and gets some acknowledgement from them (both nod, Ken also points to the screen and Alex says “yeah”) that they have





noticed the person. He also secures a response from Mary (M) who is standing across the table from him.

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 coparticipants, both in Kyoto and Tokyo. Through the technology he identifies a feature for his colleagues, they seem to recognize what is being pointed at, and Fernando assesses that they each 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.

The accomplishment of this referential activity through t-Room would seem to resonate with how such activities are accomplished in naturalistic settings. Indeed, in this case Fernando is referring to a moving image, the “referent” of his gesture disappears almost as it reaches its full extent. Fernando seems to adjust his pointing in light of not only the changing image, pulling his finger in while still holding his hand out to the place where the feature appeared, but also with regard to the conduct of both his copresent and his remote colleagues. Once he secures a response from all his colleagues, Fernando reorients away from the screen. T-Room allows, within the articulation of a single turn of talk, for participants to identify features for colleagues, to monitor their ongoing response, and transform their activities in light of the conduct of their colleagues. The accomplishment of this referential activity relies on coparticipation of others. In this case, the recipients respond in distinct ways.

The participants did engage in discussions over competing interpretations of the images, for example, regarding who the characters were, when certain features were visible, the ordering of activities, and why certain actors carried out particular actions. They also drew on what could be seen in different shots to assemble a coherent sense of a scene, to map out the sequences of actions undertaken by actors, and even to identify anomalies in what was being presented. So, for example, in the following fragment John (J), in Kyoto, is describing a sequence of activities performed by one of the characters in the clip they are currently replaying.

Fragment 4

Penny	John
4.1	4.2
	
J: (.) she looks	at him
4.3	4.4
	
J: and then (.) there	and back
P: yeah	

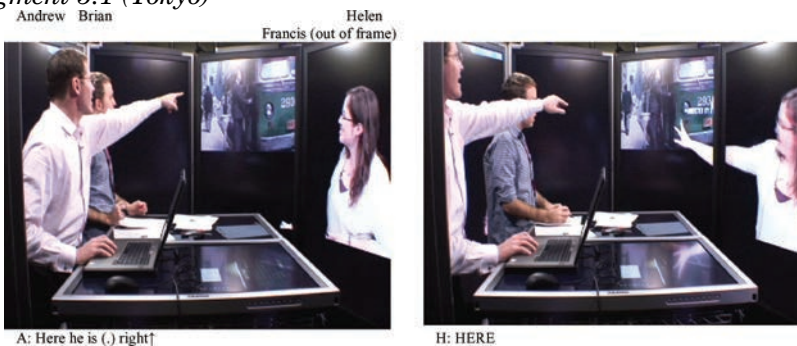
John accompanies his talk with a series of gestures to the screen. As he extends his arm and says “she looks” (1.1), Penny (P), in Tokyo, also turns toward the screen. John then provides an extended account of the clip, pointing next to a character on the right as he says “at him” (1.2), another on the left, uttering “and then there” (1.3) and back to the right again, saying “and back” (1.4). He then withdraws his hand. John’s gestures seem sensitive to the conduct of Penny, only starting the extended account once she has oriented to the screen. Penny in turn seems to follow John’s conduct, confirming part of his description. John’s conduct over the screen is shaped in light of the participation of his colleagues. He not only identifies an object but animates a series of actions that is about to occur in the clip. This relies on others seeing his action with respect to what appears on the screen, and him seeing their conduct as tied to his own and his own environment. This being achieved when the participants are actually in locations many miles apart.

The technology, therefore, seems to facilitate quite rich forms of collaboration where participants not only make use of how their actions appear to a colleague but also how this appearance can be reconfigured in the course of an activity. The participants seem to rely on their ability to produce sequentially embodied actions, to monitor the actions of the colleagues as they were being produced, and develop their conduct in light of the participation of these colleagues. Perhaps more critically, the orientation of the “shared screen” and overlaying of images of the participants onto its contents allow for a coherent way of embedding action within the environment. A participant can not only assess a remote colleague’s activities in light of his own, but can also assume that the ways in which it is being seen remotely is similar to how it is being produced locally. This is accomplished without providing participants with images of how their conduct appears in the other domain (refer to the vanity monitors often offered in video conferencing systems).

This is also despite the t-Room transforming both the appearance and timing of a participant’s conduct. As the appearance of person’s image is two-dimensional on the monoliths, a remote participant may appear to be looking at one party, while actually oriented to a local colleague standing next to him, the so-called “Mona Lisa” effect.

Moreover, time delays can mean that an action by a remote party may appear some time after it was produced. For example, in the following fragment the participants are trying to find Hitchcock in the opening scene of *North-by-Northwest*. In Tokyo, Andrew spots Hitchcock in the clip and points toward him.

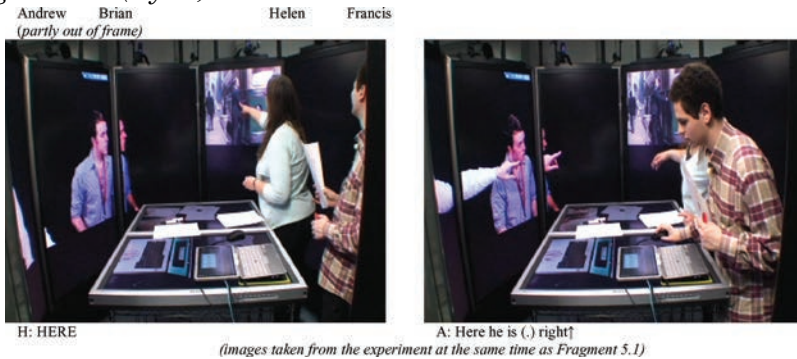
### Fragment 5.1 (Tokyo)



As Andrew brings his hand down he sees Helen also point to the screen at Hitchcock, seemingly in response to his prior action. Helen seems to confirm his identification, while also saying “here.” However, for Helen the appearance of the conduct is slightly

different. For her (in Fragment 5.2), she also points to the image of Hitchcock as he appears on the screen.

### Fragment 5.2 (Kyoto)



In Kyoto, Andrew's arm starts to extend while Helen is withdrawing her own hand back down towards the table. Hence, for Helen, Andrew's conduct appears to be responsive to her own.

In the experiments there are numerous occasions when this apparent anomaly occurs, and yet it seemed to pass unnoticed by all the participants. Indeed participants found it quite surprising when the delay was mentioned in the post hoc interviews. The participants seemed to assume a coherence to their conduct. In part, despite the delay, it is still possible to assume a coherent sequential production of the conduct; the delay of 0.3–0.4 seconds appears long enough for participants' conduct not to appear to overlap and not too long to seem noticeably absent. From what they have available and the talk and visible conduct of their colleagues, the participants seemed to have the resources not only to make sense of their colleague's actions but to assume that their own conduct was being seen as coherent. This is despite the technology introducing transformations into the spatial and temporal organization of the activity.

## 7. DISCUSSION

Technologies that aim to offer rich forms of real-time distributed collaborative work seem to offer the potential of addressing problems faced by contemporary dispersed organizations, whether these are to enhance communication between participants in different continents or to support cooperation when undertaking tasks and activities remotely. Sophisticated techniques have been developed so that participants can see presentations, “embodiments”, of their remote colleagues at a scale and orientation that replicates the effect being in the same domain. Great attention has been paid to gaze direction and shifts in orientation so that these, too, seem consistent in local and remote spaces. However, what still seems problematic, even in advanced blended systems, is how other features of work and the material environment can be integrated into these technologies. It still seems difficult to develop ways in which such technologies can help resolve those persistent and pervasive problems where participants need to refer to objects in a remote domain for colleagues. Even when great care is taken over the presentation of bodily orientation, a gesture, such as when someone simply points at an object, becomes fractured between different displays, strata, or spaces. It seems hard to provide distributed environments where participants can embed occasional features of the environment within ongoing courses of activity and interaction.

T-Room is a prototype technology that aims to provide a coherent environment in which remote participants, through their talk and visual conduct, can refer to objects and details of these objects even though they may be a great distance apart.

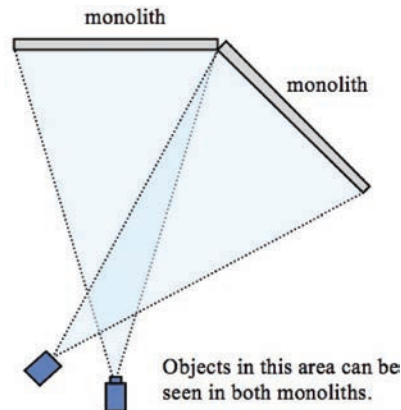


Fig. 4. For pairs of monoliths the cameras corresponding to the display can have some overlap. This overlap is greater the further conduct is away from the monoliths. Compare with similar mapping of overlaps in HP's Halo [Gorzynski et al. 2009] and "dead zones" in BISi [Paay et al. 2011]).

Participants in the experiments with t-Room did seem to manage to identify objects, animate features of these objects, and discuss fine details with both local and remote colleagues. Perhaps, more consequentially, the participants seemed able to assess the ongoing production of a colleague's activities in light of their own, so that, for example, they could recognize that a feature of the environment had been noticed and noticed in an appropriate way. T-Room did, to some extent, seem to allow for participants in the experiments to embed their activities into a geographically dispersed environment.

Because of a number of pragmatic constraints at this stage t-Room can only be assessed through quasi-naturalistic experiments. The participants did not bring to bear the different kinds of experience and expertise that academic researchers could when analyzing the materials, they were provided with. Nevertheless, the participants did manage to coordinate quite complex collaborative referential activities with both local and remote colleagues. This is despite t-Room transforming the appearance and ordering of these activities.

T-room not only transforms the appearance of gaze direction but, in common with other blended spaces, it alters the appearance of conduct across the displays. The bezels at the edge of every display mean that a remote participant's conduct appears fractured when it crosses the boundaries of two monoliths. Moreover, as each monolith is flat and arranged at angles from those on either side, there are places where two of the corresponding cameras capture part of the same image (see Figure 4). Hence, the same aspect of the conduct can appear in two images at the same time (e.g., Fragment 5.2).

These transformations did not seem to disrupt the conduct of the participants in the experiments with t-Room. In some respects, some of the constraints inherent in the design of the system seemed to facilitate the accomplishment of referential activities in the distributed environment. In contrast to the blended spaces discussed by O'Hara and Gorzynski and colleagues [Gorzynski et al. 2009; O'Hara et al. 2011; Paay et al. 2011], the designers of t-Room sought to encourage activities that were close to the monoliths. The centrally placed tabletop systems constrained the participants to be close to the screen, sometimes restricting their movement, encouraging them to stand closer to the displays and hence reducing the possibility of producing conduct that appears in multiple monoliths at the same time and perhaps diminishing the effect of rendering the participant's conduct in two dimensions.



In this article we have focused on the activities that participants accomplish over and around the shared display, however, when engaging in other aspects of the tasks they did utilize paper materials and the projections made available on the tabletop. When engaged with these materials the participants adopted a range of configurations so they could not only work together over the paper and projections but also relate these activities to what was presented on the shared display. In reviewing this materials it is apparent that this often required explicit discussions on where to position themselves and about the material they had available. There were occasions when conduct did appear fragmented and was noticeable by the participants, for example, when participants in the two different sites stood on the “same” side of the space (e.g., Figure 3, rightmost image). Here, the appearance of conduct would be incongruent, as on both sides the remote participant would seem to be behind the local one. The appearance of conduct over the shared table, such as that with the hands, would be fractured from the appearance of the body in the monolith. Moreover, when participants were in this position, if one turned “towards” the other, then for their remote colleague they would appear to be turning away. In this configuration a gesture towards an object, say the movement of hands towards a piece of paper on the table, would be fractured from the appearance of the “embodiment”. In common with other media spaces, t-Room disrupted the sequential production of referential activities. This was noticeable by the participants and may be why they tended to choose to work across the table from each other. A more thorough analysis of how participants work on such heterogeneous materials across different kinds of displays may reveal the nature of these problems and also how participants manage to assemble coherence when working across the resources they have available to them.

The experiments also point to more generic issues that need to be addressed when considering interaction in technically mediated spaces. When considering the problem of deictic reference in CVEs, Wong and Gutwin [2010] note that its successful accomplishment does not necessarily rely on its accuracy of representation and, indeed, understanding “natural pointing” might offer a resource for the design of such collaborative systems [Wong and Gutwin 2010]. It may not just be made by considering rich presentations of embodied conduct, but how its emergent and interactional nature can be made apparent so that distributed participants can collaboratively accomplish referential actions. Here, resilience to certain kinds of delay and transformation in the appearance of conduct might be consequential. Gutwin and colleagues consider, again with respect to CVEs, how participants cope with different kinds of delay and note that this may be related to the granularity of the turns being taken [Gutwin et al. 2004]. Studies of the accomplishment of naturally [2000] occurring referential activities reveal how talk, visual conduct, and features of the material environment are finely interweaved (e.g. Goodwin [2003] and Hindmarsh and Heath [1999]). When considering the effect of delays in more complex technically mediated environments not only do we need to consider how these might disrupt sequences of talk (see Ruhleder and Jordan [2001]), but also how they might transform the sequential production of visible and audible conduct. As O’Hara notes, despite the great efforts made in developing technologies like blended spaces and the care taken in trying to present “embodiments” of remote participants in coherent ways, we still have few empirical studies of how participants accomplish activities within them [O’Hara et al. 2011]. Studies of interactions in blended spaces such as t-Room might not only suggest the kinds of ways resources could be configured but also suggest ways in which we might develop novel ways of understanding and analyzing the long-standing problem of managing delays in video-mediated interaction and how we might develop techniques sensitive to these practices.

When we consider the difficulties faced by participants in media spaces, CVEs, and other forms of technologically mediated interaction [Heath and Luff 1991; Hindmarsh

et al. 1998; Luff et al. 2003], the configuration of life-sized and high-fidelity presentation of embodiments made possible in t-Room seems to provide a more coherent environment for action. However, blended spaces such as these are large and can be difficult to configure. The experiments with t-Room can suggest particular features that smaller, more lightweight technologies might need if they are to offer practical support for real-time collaborative work. Such features might include, for example, preserving the vertical orientation of the screens, but only providing displays that present the trajectory of an upcoming referential activity and the object being referred to.

The participants in the t-Room exercise did seem to manage the incongruities in the environment, and yet did so through the course of their own actions, drawing on assumptions about their and their colleague's standpoints. Although it seems like a simple action, in a referential activity such as when a person points for another to an object, participants reveal in the course of a momentary embodied action not only their relation to their own environment, but also their sensitivity to how another can see and understand that conduct, as if they themselves were in the other's place. According to Garfinkel 1963.

"By contrast the person's assumption of the interchangeability of stand-points is meant that the person takes for granted, assumes that the other person takes for granted, assumes that the other person does the same, and assumes that as he assumes for the other the other assumes for him, that if they were to change places so that the other person's here and now became his, and his became the other person's, that the person would see events in the typical way as does the other person, and the other person would see them in the same typical way he does."

Garfinkel [1963]

This assumption of the interchangeability of standpoints rests not only on considerations of the orientation of others and their viewpoint, but also on how and when others act in the environment.

Providing resources for participants to unproblematically refer to remote objects has presented challenges to those designing collaborative systems. Early media spaces took a straightforward approach to resolving these problems. They provided visible presentations of a remote coparticipant: real-time, highfidelity embodiments. These technologies allowed participants to see their remote colleagues, their facial expressions and gestures. Recent blended environments have provided high-fidelity embodiments and paid great attention to the orientation of participants' gaze and body direction. By focusing on head-and-shoulders views of participants as they sit at a desk, these efforts have addressed many of the discontinuities found in earlier media spaces and simpler commercial video conferencing systems. However, focusing on the body and embodiment presents the danger of neglecting ways in which referential activity is accomplished within a course of emerging collaborative activities.

By focusing on the body, even advanced blended spaces seem to direct attention away from the environment in which the activities are embedded. It is not just that a gesture needs to be seen accurately by a remote participant. The remote participant needs to recognize shifts in another orientation and assess the relevance of such transitions. The environment, the local ecology in which the action is embedded, is critical in this regard. A coparticipant needs to be able to make sense of a course of action as it is produced with respect to the occasioned features of the environment, and display this understanding in respect to these self-same features. When an emerging course of action appears embedded in the environment, participants do seem to be able to assemble and produce coherent coordinated conduct. When the emerging conduct is fractured within the local ecology, even if its appearance is accurate, then referential activity can be problematic.

Despite helping to direct attention towards critical aspects of everyday action, such as carefully considering the material and physical nature of conduct, the concern with embodied action may unwittingly direct attention away from how conduct is embedded within a local ecology. Inadvertently it can resonate with conceptions of interpersonal conduct offered by researchers who are concerned with how gaze direction, visual, and bodily conduct feature as communicative acts that are, despite offering sophisticated analyses of body orientation, configurations, and gestures, often considered distinct from the environments in which they are produced (e.g. Kendon [1990]; Argyle [1976], and Kendon [2004]). In order to inform the design of technologies for embodied interaction, our analyses, as Goffman [1964] suggests, need to direct attention towards how participants embed their actions within the local and material environment:

“The individual gestures with the immediate environment, not only with his body, and so we must introduce this environment in some systematic way . . . while the substratum of a gesture derives from the maker’s body, the form of the gesture can be intimately determined by the microecological orbit in which the speaker finds himself. To describe the gesture, let alone uncover its meaning, we . . . have to introduce the human and material setting in which the gesture is made” . . .

[Goffman 1964, page 164]

## ACKNOWLEDGMENTS

The authors would like to thank the participants in the t-Room experiments and the classicists involved in our study. In particular we would like to thank Alan Bowman, Mike Brady, Charles Crowther, Roger Tomlin, Melissa Terras and Segolene Tarte of the VRE for the Study of Ancient Documents (VRE-SDM) and the e-Science and Ancient Documents (eSAD) projects. We are also extremely grateful for other members of the t-Room team for supporting the experiments and members of the WIT Research Centre for helpful contributions to the analysis of the materials.

## REFERENCES

- ANDERSON, R., HUGHES, J., AND, SHARROCK, W. 1989. *Working for Profit*. Avebury, Farnborough.
- ARGYLE, M. AND COOK, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press.
- BENFORD, S., GREENHALGH, C., CRAVEN, M., WALKER, G., REGAN, T., MORPHETT, J., AND J., B. 1999. Broadcasting on-line interaction as inhabited television. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '99)*. Kluwer Academic Publishers, 179–198.
- BROUGHTON, M., PAAY, J., KJELDSKOV, J., O'HARA, K., LI, J., PHILLIPS, M., AND RITTENBRUCH, M. 2009. Being here: Designing for distributed hands-on collaboration in blended interaction spaces. In *Proceedings of Australian Conference on Computer-Human Interaction (OZCHI '09)*. 73-80.
- DE LA FLOR, G., LUFF, P., JIROTKA, M., KIRKHAM, R., PYBUS, J., AND CARUSI, A. 2010. The case of the disappearing ox: Seeing through digital images to an analysis of ancient texts. In *Proceedings of the Conference on Computer Human Interaction (CHI '10)*. ACM Press, New York, 473–482.
- DE LA FLOR, G., JIROTKA, M., LUFF, P., PYBUS, J., AND KIRKHAM, R. 2010. Transforming scholarly practice: Embedding technological interventions to support the collaborative analysis of ancient texts. *J. Comput. Supported Coop. Work* 19, 309–334.
- DOURISH, P. 2001. *Where the Action is: The Foundations of Embodied Interaction*. MIT Press, Cambridge, MA.
- DUTTON, W. H. AND JEFFREYS, P. W. 2010. *World Wide Research: Reshaping the Sciences and Humanities*. MIT Press, Cambridge, MA.
- EDEN, G. AND JIROTKA, M. 2012. Digital images of medieval music documents: Transforming research processes and knowledge production in musicology. In *Proceedings of the 45<sup>th</sup> Hawaii International Conference on System Sciences (HICSS)*. 1646–1655.
- FILIPPI, G. AND THEUREAU, J. 1993. Analysing cooperative work in an urban traffic control room for the design of a coordination support system. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '93)*. Kluwer Academic Publishers, 171–186.
- FRASER, M., GLOVER, T., VAGHI, I., BENFORD, S., GREENHALGH, C., HINDMARSH, J., AND HEATH, C. 2000. Revealing the reality of collaborative virtual reality. In *Proceedings of ACM Conference on Collaborative Virtual Environments (CVE '00)*. ACM Press, New York, 29–37.

- FUSSELL, S. R., SETLOCK, L. D., YANG, J., OU, J., MAUER, E., AND KRAMER, A. D. I. 2004. Gestures over video streams to support remote collaboration on physical tasks. *Hum. Comput. Interact.* 19, 273–309.
- GARFINKEL, H. 1963. A conception of and experiments with trust as a condition of stable concerted actions. In *Motivation and Social Interaction*, O. J. Harvey, Ed., Ronald Press.
- GARFINKEL, H. 1967. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ.
- GAVER, W. W., MORAN, T., MACLEAN, A., LOVSTRAND, L., DOURISH, P., CARTER, K. A., AND BUXTON, W. 1992. Realizing a video environment: EuroPARC's RAVE system. In *Proceedings of the Conference on Computer Human Interaction (CHI '92)*. ACM Press, New York, 27–35.
- GAVER, W. W., SELLEN, A., HEATH, C. C. AND LUFF, P. 1993. One is not enough: Multiple views in a media space. In *Proceedings of the Conference on Computer Human Interaction (INTERCHI '93)*. ACM Press, New York, 335–341.
- GAVER, W. W., SMETS, G. AND OVERBEEKE, K. 1995. A virtual window on media space. In *Proceedings of the Conference on Computer Human Interaction (CHI '95)*. ACM Press, New York, 257–264.
- GOFFMAN, E. 1964. The neglected situation. *Amer. Anthropol.* 6, 133–136.
- GOODWIN, C. 2003. Pointing as a situated practice. In *Pointing: Where Language, Culture and Cognition Meet*, S. Kita, Ed., Lawrence Erlbaum, Mahwah, NJ, 217–241.
- GOODWIN, C. AND GOODWIN, M. H. 1996. Seeing as a situated activity: Formulating planes. In *Cognition and Communication at Work*, Y. Engeström and D. Middleton, Eds., Cambridge University Press, 61–95.
- GORZYNSKI, M., DEROCHE, M., AND MITCHELL, A. S. 2009. The halo B2B studio. In *Media Space 20 + Years of Mediated Life*, S. Harrison, Ed., Springer, 357–368.
- GREATBATCH, D., HEATH, C. C., LUFF, P., AND CAMPION, P. 1995. Conversation analysis: Human computer interaction and the general practice consultation. In *Perspectives on HCI: Diverse Approaches*, A. Monk and G.N. Gilbert, Eds., Academic Press, London, 199–222.
- GUTWIN, C., BENFORD, S., DYCK, J., FRASER, M., VAGHI, I., AND GREENHALGH, C. 2004. Revealing delay in collaborative environments. In *Proceedings of the Conference on Computer Human Interaction (CHI '04)*. ACM Press, New York, 503–510.
- HARPER, R. AND HUGHES, J. 1993. What a f-ing system! Send 'em all to the same place and then expect us to stop 'em hitting: Making technology work in air traffic control. In *Technology in Working Order*, G. Button, Ed., Routledge, London, 127–144.
- HARRISON, S. 2009. *Media Space 20 + Years of Mediated Life*. Springer.
- HEATH, C. C., JIROTKA, M., LUFF, P., AND HINDMARSH, J. 1994. Unpacking collaboration: The interactional organisation of trading in a city dealing room. *J. Comput. Supported Coop. Work* 3, 147–165.
- HEATH, C. C. AND LUFF, P. 1991. Disembodied conduct: Communication through video in a multi-media office environment. In *Proceedings of the Conference on Computer Human Interaction (CHI '91)*. ACM Press, New York, 99–103.
- HEATH, C. C. AND LUFF, P. 1992a. Collaboration and control: Crisis management and multimedia technology in london underground line control rooms. *J. Comput. Supported Coop. Work* 1, 69–94.
- HEATH, C. C. AND LUFF, P. 1992b. Media space and communicative asymmetries: Preliminary observations of video mediated interaction. *Hum.-Comput. Interact.* 7, 315–346.
- HINDMARSH, J., FRASER, M., HEATH, C., AND BENFORD, S. 2000. Object-Focused interaction in collaborative virtual environments. *ACM Trans. Comput.-Hum. Interact.* 7, 477–509.
- HINDMARSH, J., FRASER, M., HEATH, C.C., BENFORD, S. AND GREENHALGH, C. 1998. Fragmented interaction: Establishing mutual orientation in virtual environments. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '98)*. ACM Press, New York, 217–226.
- HINDMARSH, J. AND HEATH, C. 2000. Embodied reference: A study of deixis in workplace interaction. *J. Pragmat.* 32, 1855–1878.
- HIRATA, K., HARADA, Y., TAKADA, T., AOYAGI, S., SHIRAI, Y., YAMASHITA, N., KAJI, K., YAMATO, J., AND NAKAZAWA, K. 2008. t-Room: Next generation video communication system. In *Proceedings of the World Telecommunications Congress at IEEE Globecom (WTC '08)*. IEEE, 1–4.
- HUTCHINS, E. L. 1995. *Cognition in the Wild*. MIT Press, Cambridge, MA.
- ISHII, H. 1990. TeamWorkStation: Towards a seamless shared workspace. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '90)*. ACM Press, New York, 13–26.
- JIROTKA, M., PROCTER, R., HARTSWOOD, M. R. S., SIMPSON, A., COOPMAN, C., HINDS, C., AND VOSS, A. 2005. Collaboration and trust in healthcare innovation: The eDiaMoND case study. *J. Comput. Support. Coop. Work* 14.
- KENDON, A. 1990. Conducting interaction. *Studies in the Behaviour of Social Interaction*. Cambridge University Press.

- KENDON, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.
- KIRK, D., RODDEN, T., AND CRABTREE, A. 2005. Ways of the hand. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '05)*. Kluwer Academic Publishers.
- KOSCHMANN, T. AND LEBARON, C. 2003. Reconsidering common ground: Examining clark's contribution theory in the OR. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '03)*. Kluwer Academic Publishers, 81–98.
- KUZUOKA, H., KOSUGE, T., AND TANAKA, M. 1994. GestureCam: A video communication system for sympathetic remote collaboration. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '94)*. ACM Press, New York, 35–44.
- KUZUOKA, H., OYAMA, S., YAMAZAKI, K. AND SUZUKI, K. 2000. GestureMan: A mobile robot that embodies a remote instructor's actions. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM Press, New York, 155–162.
- KUZUOKA, H., YAMASHITA, J., YAMAZAKI, K., AND YAMAZAKI, A. A. 1999. Agora: A remote collaboration system that enables mutual monitoring. In *Proceedings of the Conference on Computer Human Interaction: Extended Abstracts (CHI '99)*. ACM Press, New York, 190–191.
- LUFF, P., HEATH, C., KUZUOKA, H., HINDMARSH, J., YAMAZAKI, K., AND OYAMA, S. 2003. Fractured ecologies: Creating environments for collaboration. *Human Comput. Interact. J.* 18, 51–84.
- LUFF, P., HEATH, C., KUZUOKA, H., YAMAZAKI, K., AND YAMASHITA, J. 2006. Handling documents and discriminating objects in hybrid spaces. In *Proceedings of the Conference on Computer Human Interaction (CHI '06)*. ACM Press, New York, 561–570.
- LUFF, P., HEATH, C., NORRIE, M., SIGNER, B., AND HERDMAN, P. 2004. Only touching the surface: Creating affinities between digital content and paper. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM Press, New York, 523–532.
- LUFF, P. AND HEATH, C. C. 1998. Mobility in collaboration. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '98)*. ACM Press, New York, 305–314.
- LUFF, P., HEATH, C. C., AND GREATBATCH, D. 1992. Tasks-In-Interaction: Paper and screen based documentation in collaborative activity. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '92)*. ACM Press, New York, 163–170.
- LUFF, P., YAMASHITA, N., KUZUOKA, H., AND HEATH, C. C. 2011. Hands on hitchcock: Embodied reference to a moving scene. In *Proceedings of the Conference on Computer Human Interaction (CHI '11)*. ACM Press, New York, 43–52.
- MONDADA, L. 2001. Operating together through video conferencing. In *Orders of Ordinary Action*, S. Hester and D. Francis, Eds., Manchester Metropolitan University Press.
- O'HARA, K., KJELDSEKOV, J., AND PAAY, J. 2011. Blended interaction spaces for distributed team collaboration. *ACM Trans. Comput. Hum. Interact.* 18, 3–3.
- PAAY, J., KJELDSEKOV, J., AND O'HARA, K. 2011. BISI: A blended interaction space. In *Proceedings of the Conference on Computer Human Interaction (CHI '11)*. ACM Press, New York, 185–200.
- ROBERTSON, T. 1997. Cooperative work and lived cognition: A taxonomy of embodied actions. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '97)*. Kluwer Academic Publishers, 205–220.
- RUHLEDER, K. AND JORDAN, B. 2001. Co-Constructing non-mutual realities: Delay-Generated trouble in distributed interaction. *J. Comput. Support. Coop. Work* 10, 113–138.
- SACKS, H. 1992. *Lectures in Conversation: Vols. I and II*. Blackwell, Oxford, UK.
- SANCHEZ SVENSSON, M., HEATH, C. C., AND LUFF, P. 2007. Instrumental action: The timely exchange of implements during surgical operations. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '07)*. Kluwer Academic Publishers, 41–60.
- SUCHMAN, L. 2000. Embodied practices of engineering work. *Mind Culture Activity* 7, 4–18.
- TANG, J. C. AND MINNEMAN, S. L. 1991. VideoDraw: A video Interface for collaborative drawing. *ACM Trans. Inf. Syst.* 9, 170–184.
- TUTT, D. AND HINDMARSH, J. 2011. Reenactments at work: Demonstrating conduct in data sessions. *Res. Lang. Social Interact.* 44, 211–236.
- TUTT, D., HINDMARSH, J., SHAUKAT, M., AND FRASER, M. 2007. The distributed work of local action: Interaction amongst virtually collocated research teams. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '07)*. Kluwer Academic Publishers, 199–218.
- WONG, N. AND GUTWIN, C. 2010. Where are you pointing? The accuracy of deictic pointing in CVEs. In *Proceedings of the Conference on Computer Human Interaction (CHI '10)*. ACM Press, New York, 1029–1038.

- YAMAZAKI, A., YAMAZAKI, K., KUNO, Y., BURDELSKI, M., KAWASHIMA, M., AND KUZUOKA, H. 2008. Precision timing in human-robot interaction: Coordination of head movement and utterance. In *Proceedings of the Conference on Computer Human Interaction (CHI '08)*. ACM Press, New York, 131–140.
- YAMAZAKI, K., YAMAZAKI, A., KUZUOKA, H., SHINYA, O., KATO, H., SUZUKI, H., AND MIKI, H. 1999. GestureLaser and gestur laser car: Development of an embodied space to support remote instruction. In *Proceedings of the European Conference on Computer Supported Cooperative Work (ECSCW '99)*. Kluwer Academic Publishers, 239–258.

Received October 2011; revised April 2012; accepted August 2012