Embodiments and VideoArms in Mixed Presence Groupware

Anthony Tang, Carman Neustaedter and Saul Greenberg
Department of Computer Science, University of Calgary
Calgary, Alberta CANADA T2N 1N4
+1 403 220 6087

{tonyt, carman, saul}@cpsc.ucalgary.ca

ABSTRACT

Mixed Presence Groupware (MPG) is software that connects collocated and distributed collaborators together in a shared visual workspace. The problem is that collaborators in MPG focus their collaborative energies almost exclusively on their collocated partners, ignoring their distributed counterparts. This arises because remote collaborators are *disembodied* when compared to their collocated cohorts: they lack the material presence that informs others of their actions. In this paper, we recap how physical bodies facilitate collaboration in physical workspaces via feedthrough, consequential communication and gestures. We recast this theory as four design implications for *virtual embodiments* that minimize the disparity between collocated and remote collaborators within MPG. We use these properties to design *VideoArms,* a video-based mechanism that captures people's body actions within a physical workspace, and then digitally recreates them as virtual embodiments throughout the MPG workspace.

Categories and Subject Descriptors

H.5.3 **Groups and organizational interfaces**: Computer supported cooperative work.

Keywords

Mixed presence groupware, single display groupware, distributed groupware, consequential communication, embodiments, gestures.

1. INTRODUCTION

Mixed Presence Groupware (MPG) is software that connects collocated and distributed collaborators together in a shared visual workspace. In practice, we have built MPG systems by connecting several distributed displays, each supporting multiple input devices, thereby connecting both collocated and distributed collaborators [23]. Figure 1 gives a stylized example where three groups, each in a different location, work over a virtual table. Each group sees this virtual workspace on their individual displays, which can be tabletops, normal monitors, or projected surfaces. Each participant has his or her own input device, be it a finger on a touch sensitive surface, a light pen, or a mouse, and all can interact with the system simultaneously.

As a new genre of groupware [23][1][7], MPG presents not only novel technological challenges, but also subtle social issues. In particular, all MPG systems share a problem called *presence disparity*, which arises because some collaborators are physically collocated while others are remote [23]. While collocated participants are seen in full fidelity, remote participants are not. The local person can only sense the remote collaborator to the extent that the remote person is captured and presented on the computer display. This disparity hampers a collaborator's ability to pick up non-verbal cues from remote participants, but not collocated participants. Presence disparity unbalances the collaborator's subjective experience: collaborative dynamics will vary in terms of how one senses presence, engagement and involvement with their collocated *vs.* remote partners.

The core problem arises from the physical distribution of participants in the virtual workspace—the physical *presence* of collaborators varies across an MPG workgroup. In our initial informal observations of a few groups using MPG, we saw that this presence disparity has negative effects on conversational dynamics [23]. This should not have been a surprise, for observations and theories of workspace awareness [10] and video media spaces [8], when applied to MPG predict that this presence disparity would have profoundly negative effects on collaboration. Because MPG collaborators cannot communicate as effectively with remote collaborators as they can with those

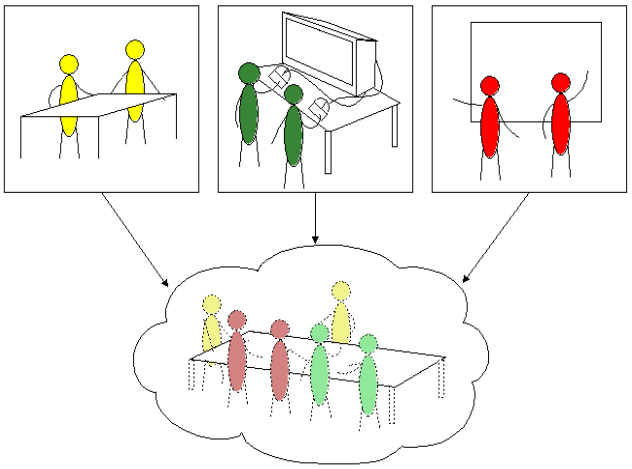

Figure 1. Three teams working in MPG over three connected displays (top), stylized as a virtual table (bottom)

Cite as:
Tang, A., Neustaedter, C. and Greenberg, S. Embodiments and VideoArms in Mixed Presence Groupware. Report 2004-741-06, Department of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4, March.

who are collocated, they will tend to focus their communicative efforts toward their collocated partners. Remote collaborators are judged less positively, are less likely to be invited into informal discussions of the work objects, and are therefore less likely to perform the task as effectively as collocated counterparts.

Our approach to mitigating presence disparity in MPG is to understand how the observable aspects of a person's presence play a role in collaboration. This virtual presence of a collaborator in a remote workspace is defined as their *embodiment.* For collocated participants, the person's entire body is the observable aspect of a person's presence. However, a body is much richer in communicative value compared to the kinds of groupware embodiments normally seen by remote participants (e.g. telepointers). To appreciate this difference, we need to understand the role a collaborator's body plays in collaborative work. We are interested in how being able to *see* a collaborator's body influences the collaboration.

This paper has two parts. First, we develop a theoretical understanding of the role embodiments have in the workspace. Using existing CSCW literature, social-psychological theories, and our own experiences with MPG, we present a set of implications for the design of MPG embodiments. Specifically, we suggest that MPG embodiments should be designed with the following properties in mind.

1. To provide feedback of what others can see, a person's embodiment should be visible not only to his distant collaborators, but also to himself and his collocated collaborators.
2. To support consequential communication for both collocated and distributed participants, people should interact through direct input mechanisms, where the remote embodiment of this input is presented at sufficient fidelity to allow collaborators to easily interpret all current actions as well as the actions leading up to them.
3. To support bodily gestures, remote embodiments should capture and display the fine-grained movements and postures of collaborators. Being able to see these gestures means people can disambiguate and interprete speech and actions.
4. To support bodily actions as they relate to the workspace context, remote embodiments should be positioned within the workspace to minimize information loss that would otherwise occur.

Second, we use these implications as a basis for understanding the shortcomings of embodiments in existing distributed groupware, and for designing effective new ones. We demonstrate that current embodiment approaches are non-ideal, and introduce a new video-based embodiment technique for MPG called VideoArms. VideoArms captures people's arms as they move over the workspace, and redisplays them as a digital overlay. As we will see, VideoArms provides feedback of what others can see, conveys rich presence information, and facilitates the richness of gestural and consequential communication in collaborative work.

2. BODIES IN COLLABORATIVE WORK

The role of the physical body in collaboration is central to understanding the imbalances in MPG interaction. In particular, we need to know what information the body gives to other collaborators, and why this information is useful. Only then can we consider what necessary features an embodiment must include if it is to correct this imbalance.

This section reviews three concepts central to bodies in collaborative work: feedback and feedthrough, consequential communication, and gestures. While these concepts are important to distributed groupware systems in general, we recast them as implications for the design of MPG embodiments.

2.1 Feedback and Feedthrough

Our ability to perceive our own bodies plays a key role in how we interact with the world [20]. We perceive our own actions and the consequences of our actions on objects as *feedback*, and we constantly readjust or modify our actions as our perceptions inform us of changes to the environment, or changes about our bodily position. Imagine threading a needle when we can see it, and compare that to the difficulty of performing the same task blindfolded. Clearly, our ability to perceive our own bodies as physical objects in the world facilitates our smooth interaction with the world.

In distributed groupware, feedback is echoed to other participants as *feedthrough*, the reflection of one person's actions on the other users' screens [4]. By observing feedthrough, remote participants can understand a person's bodily actions and the effect they have on the workspace.

Within MPG, feedback and feedthrough play a dual role. Feedback not only informs the local person of their own actions, but gives that person and their collocated partners an expectation of what feedthrough is being transmitted (and thus visible) to their remote counterparts. When feedback and feedthrough are dissimilar, this adds confusion to how local and remote participants experience the interaction.

This importance of echoing feedthrough as feedback to local participants gives our first implication for the design of MPG embodiments. *To provide feedback of what others can see, a person's embodiment should be visible not only to his distant collaborators, but also to himself and his collocated collaborators.*

2.2 Consequential Communication

Our bodies are the key source of information comprising *consequential communication*: the information unintentionally generated as a consequence of an individual's activities in the workspace, and how it is perceived and interpreted by an observer [22]. A person's activity in the workspace naturally generates rich and timely information that is often relevant to collaboration. For instance, the way a worker is positioned in the workspace and the kinds of tools or artifacts he is holding or using tells others about that individual's current and future work activities.

We see evidence of consequential communication in a wide variety of literature. Segal [22], in his studies of flight teams, found that pilots spent 60% of their time simply observing another pilot's console while it was being manipulated. Further, he reports that pilots would often react to another's actions without explicit verbal cuing. These reactions are neither overt nor awkward; rather, the graceful choreography of teamwork arises from the subtle role played by consequential communication. In small group design activities, Gutwin [10] observed that "participants would regularly turn their heads to watch their partners work." Tang's [24] reports of choreographed

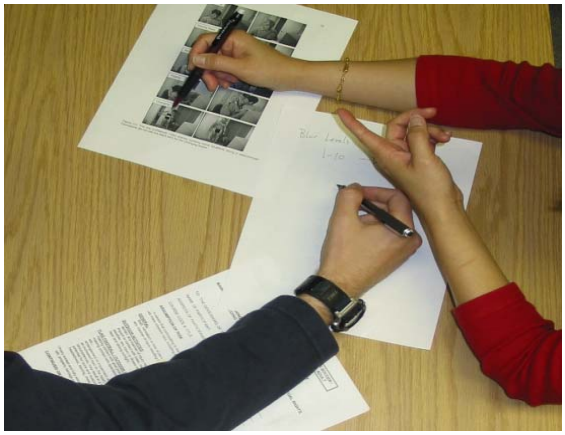

Figure 2. Corporeal arms in a common workspace.

hand movements can also be understood in terms of consequential communication: by observing others' actions and activities in a shared workspace, one can fairly accurately predict others' future acts or intentions, thereby easily working with or around them. Consequential communication is an important conduit for maintaining awareness of others, allowing us to monitor, understand and predict others' actions in the workspace without explicit action on their part [17]. Figure 2 illustrates this: the position of people's arms, how they relate to each other and the workspace artifacts, and how they are poised to do work tells a rich story of collaborators' presence, engagement and activities.

Consequential communication in MPG fails if people do not have a balanced view of their collocated and remote participants. Physical workspaces allow us to observe individual atomic-level interactions with the workspace (e.g., moving an arm towards a pair of scissors, fingers grappling at the holes of scissors, lifting and grasping the scissors, moving the scissors in hand), allowing us to predict future activities extremely well. In a virtual setting, our ability to observe others depends directly on the fidelity of the embodiment. Virtual environments typically tend away from atomic-level interactions, often preferring to represent activities at a coarser level (e.g., the mouse pointer changes into a pair of scissors, or scissors suddenly appear in the empty avatar's hand). This abruptness makes remote participants' actions less predictable. As well, computer environments for face to face work that supply indirect input devices (e.g., mice, function keys for invoking actions) can restrict consequential communication between collocated participants since they can no longer see how bodies are attached to actions, or how actions are generated [11].

One solution to this disparity in MPG is to increase the fidelity of the embodiment representation, which in turn should increase the richness of the consequential communication that is produced. Yet this also means that the system must capture appropriate information to generate a rich embodiment, which is directly related to the input mechanism of the system and how this input is connected to bodily actions. For instance, it is far more informative to observe a collaborator physically reaching over to touch and mark up a picture (on a tabletop such as in Figure 2 or on a touch sensitive surface) than to watch her cursor embodiment in the virtual workspace move over the picture via mouse input. Because her entire body is involved, it is easier to understand that *she* is the person responsible for the action in the workspace. Furthermore, in the moments prior to her touching the picture, because her whole body is moving toward the picture, it is easier for us to predict her future actions. In contrast, the cursor embodiment loses information: we do not see who it belongs to (although it could be labeled), we do not see her reach for the mouse, nor do we see her raise her finger before a button press, nor do we see where she moves to after she lets go.

If collaborators are to successfully maintain an awareness of distributed participants in MPG workspaces, then their embodiments need to be capable of providing a comparable fidelity and range of expressiveness as physical bodies. Similarly, if we are to mitigate presence disparity, we need to recognize that collocated collaborators will unconsciously use all available consequential acts to communicate ideas with one another; when distributed collaborators cannot see these consequential acts, the entire group's effectiveness suffers.

This brings us to our second implication for the design of MPG embodiments. *To support consequential communication for both collocated and distributed participants, people should interact through direct input mechanisms, where the remote embodiment of this input is presented at sufficient fidelity to allow collaborators to easily interpret all current actions as well as the actions leading up to them.*

2.3 Gestures

While consequential communications are unintentional body acts, *gestures* are intentional bodily movements and postures used for communicative purpose. Gestures play an important role in facilitating collaboration by providing participants with a means to express their thoughts and ideas both spatially and kinetically, reinforcing what is being done in the workspace and what is being said. Gestures are a frequent consequence of how bodies are used in collaborative activity: Tang [24] observed that 35% of hand activities in a physical workspace were gestures intended to engage attention and express ideas. Because intentional gesturing is so frequent, hindering the process—by not giving participants the ability to view or to produce gestures effectively—may negatively impact collaborative activities in MPG.

Two classes of gestures facilitate the communication of ideas and therefore group work in the shared workspaces: those that are purely communicative acts, and those that relate to the workspace and its artifacts. *Pure communicative gestures* arise from a person's natural communicative effort, where they can occur independently from the workspace. People use gestures to facilitate speech production [15], to emphasize parts of speech [2] and to attract attention [2]. Psychological theory suggests that spatial and kinetic gestures are part of people's semantic encoding of ideas, and therefore that the retrieval of words depend on gestures [15]. Two key pieces of evidence support this position: first, most gestures appear prior to the accompanied speech, and second, preventing speakers from using gestures tends to impede smooth speech production. For instance, Morrel, Sammuels, & Krauss [16] found that gestures usually precede speech by a 0.75s interval. More telling is that speakers' fluency has been found to be markedly hampered when they are prevented from gesturing [18]. Listeners also use accompanying gestures to interpret and disambiguate speech. Riseborough [19] found in two separate experiments that participants benefited drastically when able to view accompanying gestures compared to speech alone, both in terms of word recall and recognition. Gestures also convey semantic information above and beyond speech alone, and some replace speech entirely (e.g. yes or no via thumbs-up or thumbs-

down, insults via the middle finger). At a higher level, gestures are also used to help regulate conversation [2]. For instance, people use gestures to negotiate turn-taking (e.g., putting up your hand to express a desire to speak, or gesturing at the person who can speak next [5]).

Workspace-oriented gestures are the class of gestures that directly relate to the collaborative workspace and the artifacts within them. They typically refer to objects or locations in the workspace, or clarify verbal communication by illustration over the workspace [13]. Of course, there are many types of workspace-oriented gestures and they can be used for many different things. Bekker, Olson & Olson [2] developed a taxonomy of gestures from observations of ten different teams performing collaborative design over a workspace. First, they identified four different *types* of gestures, of which three are workspace-oriented:

- *Kinetic*: movement that illustrates an action sequence.
- *Spatial*: movement that indicates distance, location, or size.
- *Point*: fingers point at a person, object, or place. The target may be concrete, abstract, denoting an attitude, attribute, affect, direction, or location. This type of gesture is often referred to as a deictic reference.

Next, they observed that gesture types often combine into *sequences* [2]. For example, one common sequence comprising a workspace-oriented gesture is the *walkthrough:* a succession of kinetic gestures illustrating how something might be used. Another sequence is the *list,* a string of pointing gestures in concert with speech referring to a numerical or bulleted list. Sequences mean that collaborators will often combine atomic-level gestures in novel combinations to express ideas. Thus, even if an exhaustive taxonomy of atomic gestures was developed, attempting to support remote interaction by providing "canned" gestures would be an insufficient approach.

Finally, Bekker et. al. determined that gestures have several primary *roles* within a design setting [2]. The one most relevant to workspace-oriented gestures is the *design role*, where gestures relate to the current design activity and refer to things like showing distances, enacting the interaction between user and product, referring to objects, persons or places, etc. The design role is of particular interest to MPG embodiments because it emphasizes that a gesture's semantic information is often tied to the context in which it is produced. For instance, gestures in the workspace often refer to objects or locations on the workspace (e.g. "I think this object should be this big.").

Clearly, people regularly use many different kinds of both communicative and workspace-oriented gestures. Both kinds depend upon people producing gestures by animating their bodies, and upon others being able to see them in detail. This leads to our third implication for the design of MPG embodiments: *To support bodily gestures, remote embodiments should capture and display the fine-grained movement and postures of collaborators. Being able to see these gestures means people can disambiguate and interpret speech and actions*.

The above theories also confirm the importance of the relation between gestures and workspace artifacts. Yet the vast majority of distributed groupware separates the visuals of the person from the workspace. Usually the person is captured as a video stream and displayed in one window, while the workspace is shown in a different window. Even though hand gestures may be visible on the video, they are completely decoupled from the workspace. By virtue of being *about* the workspace or objects on the workspace, removing these gestures from the context of the workspace removes much of the meaning conveyed by them. Thus, our fourth implication for the design of MPG embodiments is that: *To support bodily actions as they relate to the workspace context, remote embodiments should be positioned within the workspace to minimize information loss that would otherwise occur.*

Our discussion of gestures also reinforces our second implication. Since the ability to freely use gestures is important for fluent speech production, smooth interaction in MPG is necessarily best facilitated by un-tethered input devices, where people are free to work directly over the work surface, e.g., as on a touch sensitive display. Tethering users to input devices (such as a keyboard and mouse) inhibits users from gesturing as part of their communicative effort, hindering vocabulary use and the articulation of ideas.

In summary, we have described three concepts central to how bodies contribute to collaborative work: feedback and feedthrough, consequential communication, and gesturing. We caution, however, that this list is not complete. For example, we have left out the important role of eye contact for inter-personal communication, and eye-gaze for knowing where others are focusing their attention [14][27]. However, these three concepts sufficed to help us articulate four design recommendations for embodiments in MPG.

3. EMBODIMENTS IN GROUPWARE

We now use our four design implications as a basis for understanding the most popular embodiments found in existing distributed groupware and to re-examine these approaches for their suitability within MPG.

In face-to-face situations, we watch others' bodies, their facial expressions, and the workspace to maintain workspace awareness [10][17]. In typical distributed groupware systems, we rely on embodiments to represent others [3] so that workspace awareness information can be acquired and maintained. Three approaches have dominated embodiment design in groupware: telepointers, avatars, and video embodiments. All have achieved reasonable success in distributed groupware, for they add information of varying richness where none existed before. Unlike MPG, embodiments in distributed groupware do not introduce imbalance because all collaborators see each other only through the embodiment.

3.1 Telepointers

Telepointers are the simplest approach for supporting embodiment, and were envisioned and implemented as early as 1968 [6]. Remote participants are represented in the workspace as pointers (i.e., mouse cursors), one for each person. As with the local cursor, mouse movements by participants are shown in real time as movements of corresponding pointers. This subtle visual cue is surprisingly effective in conveying a wealth of information, such as presence, location, movement, selective gestures, and activity. Telepointers also provide a reasonable estimate of where people are looking, i.e., gaze awareness [14]. Telepointers can provide implicit identity during speech, as the tight-coupling between a speaker's speech and his pointer actions means that observers are good at associating a telepointer with a speaker.

Telepointers can provide more information through judicious use of labeling, color, iconography and visual overloading [9]:

- *identity information,* where the owner's identity is made explicit, e.g., by attaching a textual name, a photo or even an abstract symbol to the cursor;
- *action information* that reflects and emphasizes its owner's actions, e.g., the rapid selection of an item is emphasized by presenting a miniature mouse with its button shown pressed as part of the cursor visuals;
- *mode information* that reflects the owner's interaction state for moded interfaces, again by changing the cursor visuals;
- *trace information,* where a visual trace of the telepointer movement over time informs where the cursor has been in the recent past [12].

In spite of their success, telepointers are limited embodiments. Remote people cannot reliably interpret an idle telepointer (has its owner stepped out for the moment, or is the owner there but not active). Telepointers provide only limited space to incorporate extra information—overloading or rapidly changing its visuals to show pointing, identity, activity information and mode can quickly make it over-cluttered and difficult to interpret. Considering Bekker et. al.'s framework [2], telepointer gestures are generally restricted to pointing; kinetic and spatial gestures are hard to perform [12].

Within MPG, the added problem is the huge discrepancy between how a person sees the telepointer of their remote collaborator *vs.* the full body actions of her collocated collaborator. The telepointer captures and presents only a fraction of body actions. It is also very small, and thus cannot create the same degree of presence when compared to a body's visual salience.

3.2 Avatars

Avatars originated in collaborative virtual environments [3], typically three-dimensional worlds with an immersive input/output system such as a CAVE, but are now mostly seen in collaborative games. Avatars often appear as humanoid, three-dimensional beings, typically with a distinctive head, body and arms. The idea is to have fairly distinctive human-like representations in what might be considered a three-dimensional simulation of the real world.

The typical collaborative avatar portrays only limited information: the location of a person in a space, and roughly where they are looking. While motion through the space is transmitted, most avatars are rendered with poor fidelity and infrequently updated, so seeing or interpreting fine-grained motion is impossible. In addition, avatars are typically abstract or pseudonymous caricatures, making identity difficult to determine.

Still, there are many versions of avatars that do portray rich information. Some allow people to animate their avatar's hand and body positions through canned gestures (see below). Gaze awareness can also be supported, by tracking where the person is looking into the virtual space and by adjusting the gaze of the avatar to point to the same direction [27]. To show identity, some systems replace the avatar head with a live video feed of its owner, thus revealing the avatar owner and facial expression (albeit at low fidelity and frame rate) [27]. People can also customize their avatars to have a more recognizable face, or to dress them with identifying clothing. Games have made significant headway in this area, providing a vast array of clothing and other bodily enhancements so that many different characters can be distinguished; however, avatars are still typically displayed with a nametag.

Still, avatars are a limited means to portray bodies in collaborative settings. While activity is carried out with the "hands" or "arms" of the avatar, only larger actions can be interpreted—the low fidelity of activity representation puts to question the utility of avatars to convey rich consequential communication. Natural gestures are fairly weakly supported by avatars, and are hampered by the poor expressiveness of their controlling devices (mice or joysticks)—typically only canned gestures are supported by keyboard-invoked waves or smiles. While data gloves and suits can fix this, they tend to be the exception rather than the rule. Finally, and as with telepointers in MPG, the fairly low fidelity avatar representation must compete with how one sees the full body of the local person—a daunting task.

3.3 Video Overlays

Many video-based teleconferencing systems use two cameras per site, one to capture a person's face, and the other to capture the workspace. People can view these two video streams usually by switching between them, or using picture-in-picture. The workspace stream lends itself to a restricted form of embodiment, since the camera captures and transmits the local person's arms as she works atop it (e.g., when the camera points down to a tabletop). The catch is that video systems like these are one-way. They do not present a shared workspace as the distant person can only see the other's workspace and interactions, but cannot work within it.

This inability to see but not interact with the distant workspace proved frustrating to several architects working within a media space system developed by Xerox PARC [28]. Their solution was to tape tracing paper atop the display of the remote workspace, and to point the local camera to this mixed paper/monitor setting. This "fused" the local and remote workspace into a single view: the camera captured the local person's arms and the marks they made on the tracing paper, as well as the remote person's arms and activities just visible through the translucent paper. Perhaps most importantly, it allowed participants to see the bodies and faces of local and remote participants *within the context* of the shared workspace.

This innovation led to several research efforts on fusing video-based workspaces. First was VideoDraw [25], a video-based solution that used multiple cameras to capture the desktop, and polarizing filters to manage video feedback (Figure 3a, top). The resulting fused image, showing both the local and remote participants' arms in the space, is exemplified in Figure 3a, bottom. As a video embodiment, VideoDraw allows a full range of fairly sophisticated gestures by giving participants a 2 ½ dimensional gesturing space (three dimensions are flattened into one, but depth cues are preserved) of each other's actions.

Using a similar technique, VideoWhiteboard [26] allowed people to draw on translucent large screens with markers, while a camera mounted behind the screen captured both the drawings and the shadows of people near the screen (Figure 3b top). These shadows were seen as silhouettes in the fused video display (Figure 3b bottom), giving the illusion that remote collaborators were on the other side of the screen. The downside is that shadows flatten the gesturing body parts to two dimensions, reducing the range of

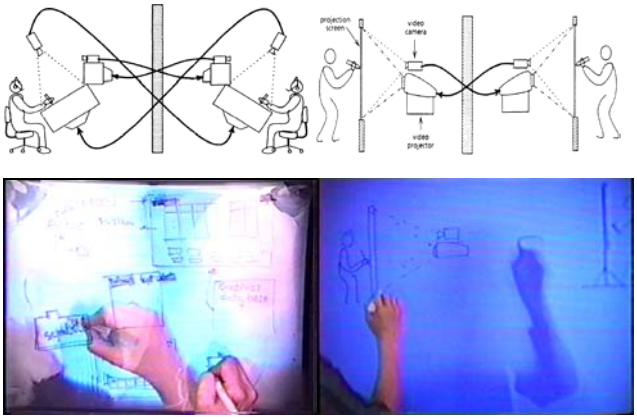

Figure 3 a) VideoDraw from [25] b)VideoWhiteboard from [26].

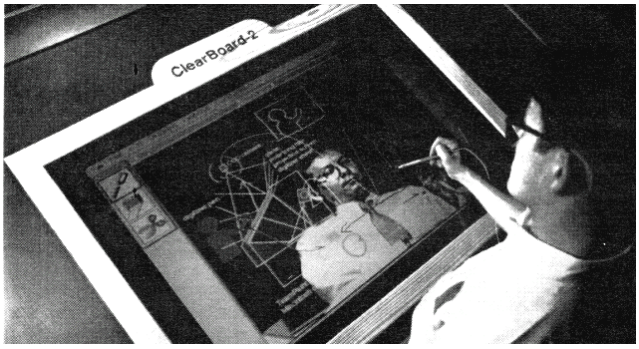

Figure 4. ClearBoard-2 from [14]

possible gestures and compromising the interpretation of detailed actions. For example, an "A-OK" sign (thumb and forefinger in a closed circle) may be seen only as a black blob because the shadow will also include the fingers behind the two front ones unless the camera angle is just right. Similarly, arm actions in front of the body may be masked by the shadow of the body itself. Shadows also hide identity, which is problematic in MPG since there is more than one person per site.

Ishii's TeamWorkstation used video-mixing technology to fuse the different video layers as overlays. Unlike VideoDraw and VideoWhiteboard, cameras could point to and fuse otherwise unrelated surfaces, such as a physical desk or a control panel. Ishii then developed ClearBoard, which used half-silvered mirrors and multiple cameras to mix a remote participant's face into the shared video workspace in a way that maintained gaze awareness [14]. As seen in Figure 4, the metaphor is of two people working on different sides of a pane of glass, where each can mark atop their side of the glass with marking pens.

While all the above systems are extremely good at capturing and transmitting live embodiments, they are limited because they are based solely on analog video technology. First, while people can see each other, they cannot manipulate the marks and artifacts created or held by others. A later version of Clearboard finesses this problem by incorporating a see-through digital display showing a groupware system and a digitizing pen for input (shown in Figure 4) [14]. This means that people can now share their electronic interactions and artifacts. Second, because these systems combine all video frames into one, they degrade substantially if more participants (and video feeds) are added.

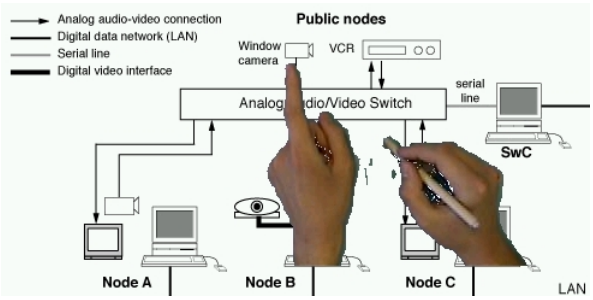

Figure 5. Using the hand as a telepointer, from [21]

To solve this, Roussel [21] has people use their arm over a solid blue surface. He extracts these arms by chroma-keying, and then super-imposes them over a digital workspace as a semi-transparent image. This gives a very crisp effect, as seen in Figure 5 (Roussel simulated the groupware capabilities, but there is no technical reason why it could not be implemented). People could also control the properties of these hands as they appeared in the workspace: their size, their relative position, and their transparency. The downside is that people must use their arms outside the workspace, although they can control what they do through the feedback that appears on the display. While reasonable for distance collaboration, such a scheme would likely be confusing to collocated collaborators in MPG.

Finally, LIDS recreates VideoWhiteboard as a digital system that enhances distributed Powerpoint presentations [1]. They capture the image of a person working in front of the shared display via consumer-grade cameras, and transform it via background subtraction and posturizing techniques into a frame containing the digital shadow. They then overlay three transparent windows to create the scene: the digital shadow, the Powerpoint frame, and a frame that captures sketching overlays. The Distributed Designer's Outpost also includes a shadowing capability captured by rear-projection [7], however the fidelity is so low that the authors state that it is useful only to indicate presence and very coarse gestures.

3.4 Discussion

Compared to the physical body, particular embodiment approaches in distributed groupware are clearly lacking in several areas, especially when applied to an MPG setting. We now discuss each of the embodiment techniques in terms of our design implications for MPG embodiments.

Our first design implication suggests that a person's embodiment should be visible not only to his distant collaborators, but also to himself and his collocated collaborators. Telepointers and avatars are typically visible by all collaborators; thus local collaborators can see how and what actions are presented to remote collaborators. Video-based embodiments are sometimes but not always visible by all collaborators (e.g., VideoWhiteboard and LIDS do not provide local feedback); the negative consequence of video feedback loops can make this hard to do in particular configurations. Still, all three approaches are potentially amenable to present MPG embodiments to both local and remote people.

Our second design implication addresses the need to support consequential communication by using direct input mechanisms and through high fidelity MPG embodiments. Telepointers perform poorly because they are typically controlled by indirect input devices, and because they presuppose a limited way for

users to interact with the system (pointing and clicking with a mouse). Thus, they present only a fraction of body actions to both local and remote participants. Avatars also fall short: most only represent activity at a coarse, high level and, excepting those controlled by data gloves or suits, also suffer from being controlled by indirect input devices. Video-based embodiments are the most promising. People use their hands and bodies to directly work within the workspace (although Roussel's does not [21]). They are able to provide rich details about the collaborators, especially when full fidelity views (vs. shadows or silhouettes) are used.

Our third design implication speaks about the necessity for embodiments to capture and display the body gestures of collaborators. The telepointer limits us too severely to adequately support all gestures, as they are restricted to motion and pointing primitives. Avatars as traditionally implemented are too coarse-grained, leaving them less than ideal (but see Vertegaal [27] and how it supports gaze awareness). Again, full-fidelity video-based embodiment approaches are the most promising, although we have to be wary of shadow-based approaches that can mask certain gestures.

Our fourth design implication stresses that embodiments should be placed within the context of the workspace. Telepointers and avatars only do this partially. While they show some actions in context, these are not connected to the owner's body that may appear in (say) a separate video window and out of context. Recall that being able to see others' bodies as they act facilitates collaboration; if the embodiment, or virtual body, has only a weak link to the physical body, then the utility of the embodiment for collaborative work is compromised. Video-based embodiments, if properly calibrated to the work surface, tightly couple the embodiment within the workspace.

In summary, we believe that video-based embodiments are the most promising approach for MPG because of their ability to capture and convey the rich gestural and consequential communication that is important in collaborative work. Yet video-based embodiments for MPG are currently problematic: analog approaches are costly (cameras, projectors and transmission bandwidth requirements), and overlaying analog video compromise scalability and image clarity. For the non-vision specialist, digital image processing has algorithm complexities and performance issues that arise during attempts to extract, manipulate and overlay high-quality images from a noisy scene. For both, the setup, and registration and calibration of equipment so that images appear in the correct place are a problem. Another problem is that the promise of video embodiments in MPG is shown by the collective properties of the various systems discussed previously, but none are designed for MPG or satisfy our implications.

4. VideoArms

To overcome some of the problems just described, we propose VideoArms as a new technique for realizing MPG video embodiments over MPG applications.

4.1 VideoArms in action

Figure 6 illustrates a snapshot of a sample session, and we will use the images to explain how VideoArms work. The top images show two connected groups of collaborators. Each group works over a touch-sensitive surface—the left is a front-projected touch-sensitive SmartBoard oriented horizontally, while the right is a rear-projected vertical SmartBoard. The surface displays a custom MPG groupware application that lets people sketch and manipulate images, while displaying video embodiments. The bottom set of images are screen grabs that reproduce what these groups see on the shared display. Not shown are cameras situated in front of the displays.

The figure illustrates what participants can see and do. First, collocated people see their own arms as local feedback. These are rendered as semi-transparent, shadow-like images, providing feedback of what others can see while minimizing distraction. Second, each group sees the solid arms of the remote participants in reasonable 2½ dimensional fidelity. Third, this is an MPG setting, where all can gesture simultaneously[1]. Fourth, both physical and video arms are synchronized to work with the underlying groupware application, where gestures and actions all appear in the correct location. Fifth, arms preserve the physical body positioning relative to the workspace. For example, because the people at the table display are standing at the back side of the image, their arms appear on the vertical display as coming from the top. Sixth, participants use untethered direct input methods, such as touch and pens, to interact with the groupware application.

From a collaborative standpoint, the VideoArms prototype satisfies our MPG embodiment requirements.

1. Local participants know what remote people see because feedthrough is shown as feedback.
2. Consequential communication of actions is high because the body is the input device to the touch sensitive surface. Other collaborators can easily predict, understand and interpret another's actions in the workspace as one reaches towards artifacts and begins actions.
3. Rich gestures (coupled with conversation and artifact manipulation) are also high because the display of the embodiment is of good 2½ d fidelity and of sufficient frame rate (~15-20 fps). Task-related gestures work because arm position is related to the objects in the workspace.
4. Collocated participants can use and interpret natural body language of their physical bodies as they communicate and work. Because all actions are untethered, direct and in the workspace context, the individual's physical body *is* the embodiment.

Our theory-based design bears out in practice. Using a puzzle-based task, we ran an observational pilot study to evaluate the feasibility of VideoArms as an embodiment technique. Participants reported that they liked the ability to see their remote collaborator work. We saw them use their dominant hand to interact with the groupware application, and use both hands to gesture to the other collaborators. We also saw evidence that participants used consequential communication to predict others' actions and modify their behavior accordingly. For example, we saw participants "back off" a workspace artifact if their collaborators were reaching for it.

[1] Our underlying groupware application currently allows only one person per site to interact with it, a temporary limitation arising from the touch display technology we are using. We previously implemented several true MPG systems that allow multiple interactions per site [23], and are now making them work with Smart Technologies' DViT touch displays.

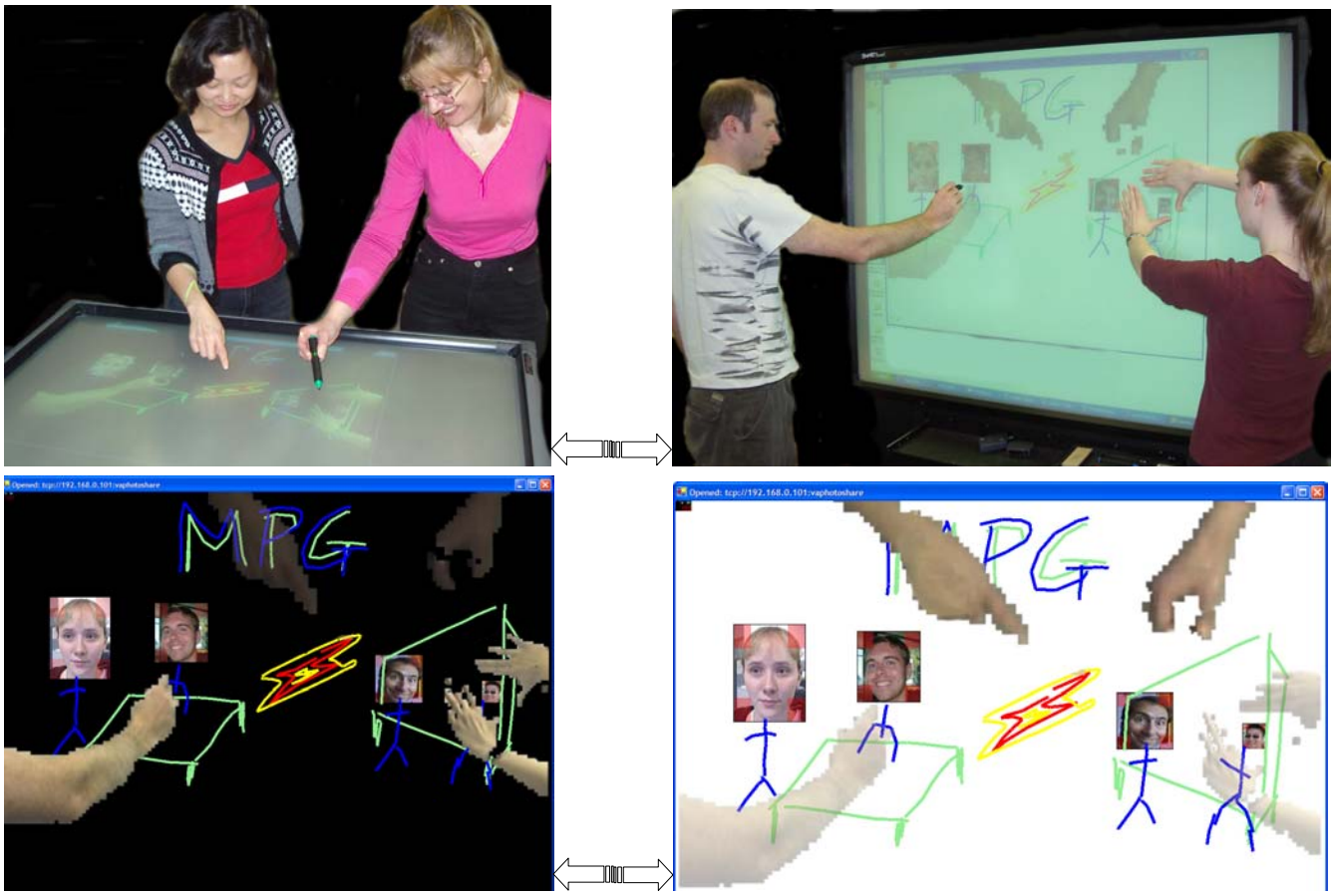

Figure 6. VideoArms in action, showing two groups of two people working over two connected MPG displays (top) and a screen grab of what each side sees (bottom). Local and remote video arms are in all scenes, but local feedback is more transparent.

4.2 Implementation

VideoArms uses inexpensive web cameras positioned approximately 2 meters in front of the display (actual position depends on the display size). The software extracts the arms (and other bare-skinned body parts) of collaborators as they work directly over the displayed groupware application. It transmits these images to the remote workstation, where they are further processed to appear as an overlay atop the remote workspace. For local feedback, it also captures and overlays a local person's video arms over the local work surface.

Frames captured by the camera are processed, transmitted and displayed in a four step process.

The first step finds the regions in the video frame that match skin color. Images are converted from RGB (red, green, blue) to HSV (hue, saturation, value) colourspace, which confines skin tones across race to a fairly small HSV colour region. A brute force matching algorithm determines which pixels in the image correspond to skin tones, thereby creating a skin mask. Morphological opening, a standard computer vision technique, is then applied to the skin mask to remove image noise while still preserving the shape and size of larger objects. We now have a silhouette image of the collaborator's arms similar to what is seen in Figure 7 (top left) and to the shadow-like embodiment found in other systems [7][1][26].

The second step produces real arms in full-colour and fidelity. It does this simply by overlaying the mask with pixels from the original image. An example of this image is Figure 7, top middle.

The third step transmits this image to listening clients via IP multicasting (clients include both the remote and local display). IP multicasting is used to reduce the amount of data on the network, and its use of UDP packets ensures quick delivery. Of course, other networking techniques could be used but care has to be taken to preserve performance.

The fourth step uses standard GUI techniques to draw all received

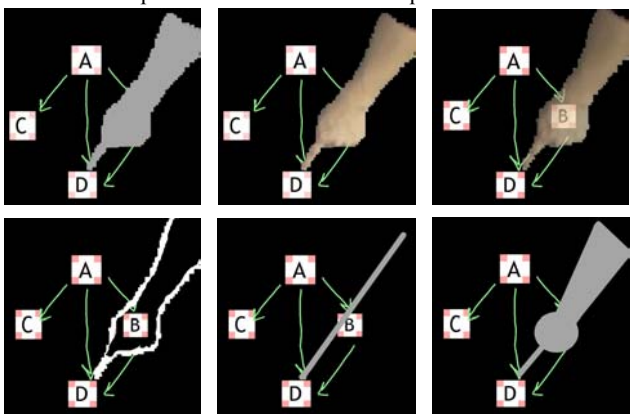

Figure 7. Various image manipulations of a video arm

images on top of the groupware work surface, which creates a composite of local and remote arms.

We should add that because our video arms are completely digital images, they can be rendered in many ways, either before or after the image is transmitted. For example, several arms in Figure 6 and the arm in Figure 7 (top right) are rendered semi-transparent. Other possible techniques include outline, vector, and stylized arm representations (Figure 7, bottom), and a means to change the size of the arm (as done by [21]). As well, unlike analog video systems, the digital nature of our VideoArms means that it can be applied to any kind of groupware system.

As seen in Figure 6, our VideoArms are not perfect. While certainly useful, they are somewhat jaggy and noisy. They also appear at roughly 15-20 frames per second rather than the 30 fps recommended in cinematography. This is because we are primarily interested in CSCW design research *vs.* computer vision research; we use only elementary and well-known image processing techniques in our prototype. Undoubtedly, true computer vision researchers could improve on our method of extracting arms from the scene while minimizing processor demands.

VideoArms is built using Python, the .NET Framework, the PyVideoCapture, Python Imaging Library, and Python numarray open source libraries. Several inexpensive cameras were used, ranging from an Intel CS430, a Logitech QuickCam Pro 4000, and a Winnov Videum camera.

To maximize performance, we use one computer to process and transmit the captured video (ideally this could be done by a special purpose hardware board), and another to display the VideoArms and run the groupware application. On a Pentium III 1.4GHz PC video frames are processed at 320x240 resolution at 15-20 fps, which is overlaid across a high resolution 1024x768 groupware workspace. This resolution is reasonable for interpreting consequential communication and gestures. It also improves upon LIDS, which works over a 640x480 workspace and a 176x144 video image on a 1GHz machine [1].

4.3 Calibration

The current VideoArms implementation is fairly robust across human skin-types, but requires two calibration steps on a per-location/per-camera basis.

First, each time a new camera is used, the system must be calibrated to understand "skin tones" from the camera. Because webcams are generally fairly low quality, their picture quality, colour range and image sharpness differ drastically from model to model. As a result, colour components (R,G,B, and their counterparts H,S,V) that register on one camera as skin may not register as skin on another and vice-versa. Typically, we use a corpus of ten images taken with the camera to determine appropriate values for skin tones for that particular camera.

Second, to correct for imperfect camera-screen alignment, a short five second calibration sequence is run when VideoArms is started. The problem arises because the camera is rarely positioned such that the groupware application perfectly fills the camera frame. A simple calibration wizard shows the camera frame, and asks the user to select three corners of the groupware application. From here, the wizard determines how much of the camera image to crop, and how the image needs to be transformed so that the composited image displays arms accurately.

4.4 Implementation nuances

As we developed VideoArms, we identified twp difficulties that were solved via workarounds. We include our solutions for those wishing to replicate and perhaps improve our implementation.

First, VideoArms performs image segmentation to determine where "skin" is in the image. This approach works well when the skin/hands are being used over a rear-projected display, plasma or CRT display. However, in front-projection systems detecting skin is more difficult: people's hands are interposed between a projector and the physical surface and the bright light of the projector shining on people's hands washes out their skin tones. To reduce skin discoloration, we limited the color palette of the front-projected groupware workspace to dark tones (e.g., black, brown and evergreen). A better solution is to predict and detect the discoloration on the skin given the particular pixel colors being projected; however, such an algorithm is not well understood and is likely computationally expensive.

With high quality video cameras and bright projection displays, the camera can capture not only people's physical bodies, but also the VideoArms projected on the workspace. This can result in visual feedback loops if the algorithm perceives the projected remote VideoArm images as skin. To solve this problem, we paint the images of remote arms slightly off-colour so that they are not captured by the system. This seems to work well in practice. LIDS also report this problem [1], but they use a more complex image-processing technique to remove shadows after they are captured.

5. CONCLUSION

Our research makes two primary contributions. First, we offer four implications for the design of Mixed Presence Groupware (MPG) embodiments, all grounded in existing CSCW research, social-psychological theories, and our own experience with MPG systems. We present an understanding of social issues in MPG systems, and in particular explain why embodiments should incorporate feedback, consequential communication and gestures to mitigate the presence disparity problem. Our recommendations give guidance to those designing MPG embodiments and technologies.

Second, we contribute VideoArms, a video-based embodiment designed around the four implications. We explained how VideoArms naturally supports feedback and feedthrough, intentional and unintentional gestures, and consequential communication over MPG groupware surfaces, thereby reducing the presence disparity problem.

VideoArms is not a total solution. For example, eye contact and body positioning, which have been found to be important to collaboration are not supported at all. Yet the VideoArms embodiment is a reasonable first step for a workspace-focused group because it presents those parts of the body within the workspace context.

VideoArms is a working proof of concept, and as such there is still room to improve its interface as well as the underlying groupware system. These need to be fixed, at which point we will undertake a thorough empirical evaluation to validate VideoArm's effectiveness as an MPG embodiment. At this point, however, we believe that we have forwarded MPG research into a space where we can begin to understand embodiment design and

the tradeoffs between different types of embodiment in MPG collaboration.

6. ACKNOWLEDGMENTS

This work was partially funded by NSERC, and equipment donations from Smart Technologies Inc. We are grateful to other members of our laboratory for intellectual contributions and feedback.

7. REFERENCES

[1] Apperley, M., McLeod, L., Masoodian, M., Paine, L., Philips, M., Rogers, B., and Thomson, K. Use of video shadow for small group interaction: Awareness on a large interactive display surface. *Proc 4th Australasian User Interface Conference (AUIC'03)*, 81-90, 2003.

[2] Bekker, M. M., Olson, J. S., and Olson, G. M. Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. *Proc ACM Designing Interactive systems (DIS'95)*, 157-166, 1995.

[3] Benford, S., Greenhalgh, C., Bowers, J., Snowdon, D., and Fahlén, L. User embodiment in collaborative virtual environments. *Proc ACM CHI '95,* 242-249, 1995.

[4] Dix, A., Finlay, J. Abowd, G. and Beale, R. *Human-Computer Interaction (2nd Edition),* Prentice Hall, 1998.

[5] Duncan. S. Some signals and rules for taking speaking turns in conversations, *Journal of Personality & Social Psychology*, 10. 283-292, 1972.

[6] Engelbart D. and English W. A research center for augmenting human intellect. *Proc Fall Joint Computing Conference,* 33, 395-410, AFIPS Press. 1968.

[7] Everitt, K. M., Klemmer, S. R., Lee, R., and Landay, J. A. Two worlds apart: Bridging the gap between physical and virtual media for distributed design collaboration. *Proc ACM CHI'03*, 553-560.

[8] Finn, K., Sellen, A. and Wilbur, S. *Video-Mediated Communication.* Lawerence Erlbaum Associates, Inc, 1997.

[9] Greenberg, S., Gutwin, C. and Roseman, M. Semantic telepointers for groupware. *Proc IEEE 6th Australian Conf on Computer-Human Interaction (OzCHI'96)*, 54-61, 1996.

[10] Gutwin, C., and Greenberg, S. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work* 11(3-4), 411-446, Kluwer, 2002.

[11] Gutwin, C. and Greenberg, S. Design for individuals, design for groups: Tradeoffs between power and workspace awareness. *ACM Proc CSCW'98,* 207-216, 1998.

[12] Gutwin, C., and Penner, R. Improving interpretation of remote gestures with telepointer traces. *ACM Proc CSCW'02,* 49-57, 2002.

[13] Harrison, S. and Minneman, S. A bike in hand: A study of 3-D objects in design. In Dorst, K, Christiaans, H. and Cross, N. (Eds), *The Delf protocols workshop: Analyzing design activity*, 205 – 218, 1994.

[14] Ishii, H. and Kobyashi, M. Integration of interpersonal space and shared workspace: Clearboard design and experiments. *ACM Transactions on Information Systems,* 11 (4), 349-375, 1993.

[15] Krauss, R., Dushay, R., Chen, Y. and Rauscher, F. The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, 31, 533-552, 1995.

[16] Morrel-Samuels, P. & Krauss, R.M. Word familiarity predicts the temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 18, 615-623, 1992.

[17] Pinelle, D., Gutwin, C. and Greenberg, S. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Trans Human Computer Interaction*, 10(4), 281-311, 2003

[18] Rauscher, F. B., Krauss, R. M., & Chen, Y. Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science,* 7, 226-231, 1996.

[19] Riseborough, M. G. Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior,* 5, 172-183, 1981.

[20] Robertson, T. Cooperative work and lived cognition: a taxonomy of embodied actions. *Proc 5th European Conference on Computer Supported Cooperative Work (ECSCW '97)*, Kluwer, 205-220,1997.

[21] Roussel, N. (2001) Exploring new uses of video with VideoSpace. *Proc 8th IFIP International Conference on Engineering for Human-Computer Interaction (EHCI'01),* LNCS 2254, 73-90, Springer.

[22] Segal, L. D. Designing team workstations: the choreography of teamwork. *Local applications of the Ecological Approach to Human-Machine Systems,* vol 2, P. Hancock, J. Flach, J. Caird and K. Vicente (eds.), LEA Press, 1995.

[23] Tang, A., Boyle, M. and Greenberg, S. Display and presence disparity in mixed presence groupware. *Proc 5th Australasian User Interface Conference (AUIC'04)*, 73-82, 2004.

[24] Tang, J. (1991). Findings from Observational Studies of Collaborative Work. Intl J Man-Machine Studies, 34(2), 143-160.

[25] Tang, J., and Minneman, S. Videodraw: A video interface for collaborative drawing. *ACM Transactions on Information Systems,* 9 (2), 170-184, 1991.

[26] Tang, J. and Minneman, S. VideoWhiteboard: Video shadows to support remote collaboration. *Proc ACM CHI'91,* 315-322, 1991.

[27] Vertegaal, R. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. *Proc ACM CHI'99,* 294-301, 1999.

[28] Xerox PARC (1987) *The Office Design Project. Systems Concept Laboratory*, Video report.