WHAT VIDEO CAN AND CAN'T DO FOR COLLABORATION: A CASE STUDY

Ellen A. Isaacs and John C. Tang
SunSoft, Inc., 2550 Garcia Ave., Mountain View, CA 94043
ellen.isaacs@sun.com; john.tang@sun.com

ABSTRACT

As multimedia becomes an integral part of collaborative systems, we must understand how to design such systems to support users' rich set of existing interaction skills, rather than requiring people to adapt to arbitrary constraints of technology-driven designs. To understand how we may make effective use of video in remote collaboration, we compared a small team's interactions through a desktop video conferencing prototype with face-to-face interactions and phone conversations. We found that, compared with audio-only, a video channel adds or improves the ability to show understanding, forecast responses, give non-verbal information, enhance verbal descriptions, manage pauses and express attitudes. These findings suggest that video may be particularly useful for handling conflict and other interaction-intense activities. But the advantages of video depend critically on the nearly-instantaneous transmission of audio, even if it means getting out of sync with the video image. On the other hand, when compared with face-to-face, it can be difficult in video interactions to notice peripheral cues, control the floor, have side conversations, point to things or manipulate real-world objects. To fully enable rich interactions, video should be integrated with other distributed tools that increase the extent and type of shared space in such a way that enables natural collaborative behaviors within those environments.

KEYWORDS: Remote collaboration, video conferencing, computer-supported cooperative work, user interfaces.

INTRODUCTION

Previous work on collaborative systems has revealed that building tools for groups of people involves specific challenges beyond those for single-user systems. Collaborative systems must be designed so that they are both useful and usable enough to induce a critical mass of people to adopt the technology [8,10]. When multimedia technology is included in collaborative systems, more design challenges are added, since so little is known about how to combine different media in ways that are natural for people to use. At the very least, we know that incorporating multimedia into a computer system requires more than just attaching video or audio onto the front end without rethinking the entire user interface [28].

There has been particular interest in the use of video to enhance remote collaboration, which has traditionally been supported by voice-only (phone) or text-only (e-mail) interactions. Although it would seem that video would greatly improve the quality of interactions among remote participants, many studies have found no evidence that groups are more effective or efficient at solving problems or making decisions when they are connected through a video and audio link than when they use only an audio link [2,9,15,20, 26].

However, we believe there is still good reason to pursue video as an integral part of collaborative technology. These previous studies measured the *product* (e.g. decisions, solutions, completion times) of interactions among strangers who were asked to accomplish an artificial task for the purposes of the study. The effects of video are more likely to be visible when studying the *process* of interactions, particularly among people who know each other and are accomplishing real work. For example, video is likely to be useful for managing the mechanics of conversations, e.g. turn taking, monitoring understanding, noting and adjusting to reactions [3,4,12,18, 26]. If video is effective at enhancing the process of interaction, at the very least it will encourage remote coworkers to collaborate more frequently. In addition, we suspect that richer interactions are likely to lead in the long run to more and/or higher quality results, although the connection in any given instance may be subtle and difficult to capture in short-term laboratory experiments. As Gale [9] notes:

> The structure of groups is continually changing. The effects of technology on a group may take weeks, months, or even years before becoming apparent. These sort of effects cannot be fully explored in a one hour experiment (p. 187).

If the process is important to collaboration, then the mechanics of interaction must be facilitated in the user interface so that users may take advantage of their rich set of existing skills in a natural and intuitive way. Video is also worth studying as a tool for collaboration because the market is driving the integration of video into many collaborative systems. We must guide the design of these new systems to make the most effective use of multimedia capabilities.

To study the user interface implication of using video for remote collaboration, we observed a team of engineers who were using a desktop video conferencing prototype (DVC) [22]. The prototype ran on Sun workstations with a prototype add-on board that enabled real-time video capture, compression, and display (combined with built-in audio capability) to bring digital audio-video conferencing onto the workstation desktop. Rather than conducting a broad survey of users' reports of their perceptions in using this technology, we focused on studying the details of one group's behavior when using video and audio as compared with audio-only and face-to-face interactions. Our intention is to describe the evidence we found for the benefit of video in remote conversations over audio alone, and to point out how video interactions fall short of, and in some ways offer advantages over, face-to-face interactions. We then discuss how our results may be applied to the design of effective video conferencing systems.

ACM Multimedia 93 /6/93/CA, USA

Published in the Proceedings of the Multimedia, 1993, ACM Press: Anaheim, CA, 199-205.

METHOD

We observed a team of five software engineers who were distributed across three sites. Two worked in a building in Billerica, MA, two worked in a building in Mountain View, CA, and one worked in another Mountain View building about 500 yards away from the first. The team had previously worked together when they were all located in Billerica, but they had recently moved to their distributed locations for reasons unrelated to this study.

At the time of this study, the prototype video board used the Intel RTV 1.5 video compression algorithm. The video windows had a video resolution of 120 x 128 pixels, although that resolution could be scaled to any arbitrary size window. The default video frame rate was five frames per second, due to some long-distance network bandwidth limitations, but the users could request a different video frame rate before starting a conference (although they did so only once). More details on the technical description of the prototype may be found in Pearl [17].

Although we took many measures of their work activity, the data for this paper are based on videotapes of six interactions: two desktop video conferences, two face-to-face interactions and two telephone conferences. (See Tang and Isaacs [23] for a further description of the study results.) One of the desktop video conference meetings included all five group members (call them Kate, Jeff, Jack, Dave and Craig) and one was between just Kate and Jeff. The two face-to-face meetings included the same sets of participants. We could not obtain phone conference data among the same sets of people, so instead we studied a four-way call between Kate, Jeff, Jack and Dave, and a three-way call between Jeff, Craig and Dave. Table 1 shows the people in each interaction we observed.

Phone	DVC	Meeting
Jeff, Craig Dave	Kate, Jeff	Kate, Jeff
Kate, Jeff, Jack, Dave	everyone	everyone

TABLE 1. People involved in each observed interaction.

The five-person DVC was actually a three-way connection where two people crowded around one camera and workstation at each of two sites. The four-person phone conference actually connected three sites; two people were in the same office sharing a telephone speakerphone.

BENEFITS OF VIDEO OVER AUDIO ONLY

A detailed analysis of the video tapes brought out the benefit of video conferencing. Specifically, participants used the visual channel to: express understanding or agreement, forecast responses, enhance verbal descriptions, give purely nonverbal information, express attitudes through posture and facial expression, and manage extended pauses.

Expressing Understanding

The most common use of the visual channel was to show understanding and, in some cases, agreement by nodding the head while someone was speaking. Research has shown that speakers are quite adept at adjusting the content of their utterances to their addressees' level of understanding [3,4,12]. Furthermore, they expect different degrees of feedback depending on the complexity of the topic [12]. Head nods are a subtle and non-intrusive way to convey understanding [5], and they were used extensively throughout the DVCs. Participants nodded their heads to different degrees and at different rates, showing different levels of understanding. Sometimes they leaned forward to indicate they were still trying to understand, and other times they looked away and tilted their heads, indicating they were considering the idea. For instance, during the two-way DVC, Kate explains a technical issue. At first, Jeff tilts his head and looks puzzled, but eventually he gives a slight head nod as he grasps the concept. Immediately after that, he sighs and shakes his head, acknowledging the issue as difficult. All these subtle reactions give Kate a running commentary on the state of Jeff's understanding. Later, Kate asks him to confirm his understanding of an idea and he says "Uh huh," but then he looks down and purses his lips as he considers the issue. Kate proceeds to elaborate, apparently responding to the visual rather than the auditory feedback.

In contrast, during the phone conferences, speakers often explicitly asked for confirmation. In one instance, Dave says, "...we should probably take, like, the first part of the meeting and just go through and see what questions you guys have." After a three second pause, he says, "Okay? Then you can at least get your questions answered. [1 sec. pause]. And then we can hit you up for stuff that we want to know. [1 sec.]. Okay? [1 sec.] All right?" Finally, Jeff says "Yep" and continues. With no visual feedback, Dave had to request a response four times before getting one.

In DVCs, the video provided an effortless and ongoing feedback channel that gave the participants a fluid sense of each other's understanding throughout the conversation. Without the video, the participants had to work harder to get much less information about each other's understanding.

Forecasting responses

In the desktop video conferences, the participants not only indicated their level of understanding, but they occasionally forecasted their response to each others' remarks through their gestures. Often they indicated their responses by shaking their heads or making facial expressions. For example, in the two-way DVC, Kate makes a point and Jeff tips his head left and right in a gesture indicating "sort of." When she finishes, he starts his turn with "Yeah, but..."

Later, Kate starts to nod in response to Jeff's comment but then stops abruptly, indicating she thought she agreed but now isn't sure. When she gives no indication of agreement at the end of his utterance, he prompts her with "Right?" He seems to ask for explicit feedback because she stopped nodding in the middle of his utterance. Forecasting negative responses was just one way that participants seemed to use the visual channel to express and handle disagreement. Others will be discussed in the examples below.

Obviously it is impossible to use head gestures and expressions to forecast responses on the telephone. As a result, participants are unable to read each others' gestures and adjust their utterances in mid-course. This is not to say that addressees don't recognize that their reactions aren't being forecast and therefore explicitly express their reactions verbally. But doing so requires more effort, and so people may be more inclined to let subtle problems pass. In particular, participants may prefer not to verbally express disagreement that might have been reflected on their faces. The speaker may therefore be unaware of a potential problem and cannot take steps to work out the disagreement.

Enhancing Verbal Descriptions with Gestures

We also observed a variety of cases when DVC participants made non-arbitrary gestures that emphasized their point. For example, Kate makes a succession of gestures during her conference with Jeff. She says, "It really helps me when I draw little diagrams [makes a drawing gesture] just to make me think of how things (unintelligible) [interlocking her fingers, shown in Figure 1]. There's so many functions now, the diagrams get all [flicks wrists back and forth showing a scattered feeling], get messy really quickly..." We cannot know whether Jeff understood the words we could not decipher, but her gesture indicates that she thinks the diagrams help her see how things *fit together*. Finally, she uses the "scattered" gesture to finish her thought and then follows it up with words. All these gestures convey shades of meaning that enhance Jeff's understanding.

FIGURE 1. Gesturing accompanying talking: A sequence of two images in time shows Kate (upper window) making a gesture to indicate *fit together*.

In many cases, the gestures appeared to be made unconsciously, sometimes outside the view of the camera or when the other person was not looking. Many people gesture while talking on the phone, apparently because it helps them express themselves verbally. As a result, when people cannot see each other, they may not express verbally the subtleties conveyed through their inadvertent gestures. Of course, these lost shades of meaning may rarely have a dramatic effect, but there are cases when they enhance participants' understanding noticeably and possibly in critical ways. Especially during an extended conversation, seeing each other's gestures is likely to increase participants general level of mutual understanding without requiring extra effort.

Conveying Purely Non-verbal Information

Not only did DVC participants use gestures to forecast their reactions and to emphasize their points, they occasionally responded solely with gestures, such as shaking or nodding their heads, shrugging, smiling, looking confused, or giving a specific meaningful gesture. For example, in the five-way DVC, Jack is frustrated about a decision, and asks "What does that benefit [this project]?" He then makes a "zero" gesture with his hand and says no more. In the two-way DVC, Kate and Jeff finish discussing a problem that they are not in a position to resolve themselves. They look at each other and make facial expressions that express "Oh well." Jeff shrugs and raises his hands, again as if to say, "c'est la vie." They then move on to the next topic. Of course they could have expressed their sentiments verbally, but this interaction highlights the ease and subtlety of interaction that video allows. It also illustrates that, in contrast to the predominantly serial nature of audio interaction, video supports concurrent interaction. Through their simultaneous gestures, they were able to realize that they both reached the same conclusion at the same time.

In another example of using visual information, Jeff notices that another person, Ted, is walking behind Craig and Dave as they are discussing a technical matter. Ted happens to be knowledgeable about the matter, so Jeff suggests asking him to join the conversation, which he does. Clearly, it would be impossible for a phone conference participant to draw someone at a remote site into the conversation; only the person on that end could do so.

Participants could not convey information purely non-verbally over the phone. One interesting incident occurred in the four-way phone conference, which occurred only because Jack and Kate were in the same location. At one point, Jeff asks Jack, "I forget, how big of a pain it is to add new built-ins, Jack?" After a three second pause Kate observes, "He doesn't look too happy," and Dave bursts out with a laugh. Had Kate not been next to Jeff, the pause would have indicated only that he was considering the answer; Jack's spontaneous unhappy expression would have been lost.

Expressing Attitudes In Posture and Facial Expression

The previous section described instances when informational content was conveyed visually. We also saw many instances in the DVCs when a person's *attitude* about verbal content was conveyed through posture and facial expression. The participants used facial expressions to indicate skepticism, surprise, amusement, confusion, conviction and so on. For example, at one point in the five-way DVC, Jack gives a treatise on an issue as he leans forward, moves his torso around and gestures with his arms. There can be no question about the strength of his conviction.

FIGURE 2. Visually demonstrating humor: Craig (upper left video window) throws head back when others smile, showing appreciation of humorous response.

Later in this conference, Jeff tells the group he has written a software utility they can use. They express interest, but then Craig teases Jeff, "As usual, no documentation." Jeff smiles and says, "It's not even done yet!" Craig throws his head back while smiling broadly, as shown in Figure 2. Jeff's words could be construed as defensive, but the smiles and Craig's response makes it clear to everyone that the conversation is in fun. In this case, the context indicated that Craig was teasing Jeff, but it is easy to imagine a situation when it would be important to confirm that one's humor was appreciated. A reaction of silence (with an unseen smile) could easily be misinterpreted.

It was particularly interesting to see how participants used visual cues to convey disagreement. In many cases, participants looked away from a speaker when they disagreed with what she was saying, sometimes returning their gaze as soon as she said something they agreed with or when the topic changed. In other cases, they responded in understated terms but looked down and sat back in their chairs while doing so. The conflict was communicated indirectly but effectively. It is impossible to convey such information through the phone. Participants must use more explicit techniques to register disagreement, which can make it more difficult to negotiate constructively.

Managing Pauses

Finally, the visual channel was particularly effective for interpreting the meaning of pauses, which can be helpful in determining someone's intention. The participants frequently interpreted pauses as indicating a lack of understanding and responded by further elaborating. However, we observed instances where the video indicated other meanings for a pause. For example, in the two-way conference, Kate responds to a question by looking to her left and consulting her notes for 13 seconds. Meanwhile, Jeff waits without trying to clarify his question. At another point, Jeff agrees to do something, and then scribbles a note to himself for the next 12 seconds. Kate looks up, sees what he's doing and waits until he is done.

The video also made it easier to manage extended pauses, which generally must be explained in phone conversations. In one dramatic example during the five-way DVC, the two Billerica participants spend over two minutes looking for an electronic mail message while the others wait. There are extremely long pauses, punctuated by the other three teasing the two in Billerica and having a casual conversation among themselves. The Mountain View participants are able to monitor the other two members' progress and adjust their expectations accordingly.

There were certainly instances of non-problematic pauses during the phone conference as well. In fact, one lasted as long as 28 seconds. However, on the whole they were more likely to be explained explicitly. At one point in a phone conversation, Dave says "I'm trying to look down things that are open bugs," meaning that he is consulting a list. For the next seven minutes, his participation in the conversation is minimal, until he says "I can't find anything else in here."

Design Implications of Adding Video

Our results clearly showed that video provides a great deal of information that participants use to enhance their interactions relative to phone conversations. People have extensive experience interpreting small changes in expressions, gestures, and body position and adapting their talk in response. The video channel enabled participants to take advantage of those cues. Our users appeared quite adept at transferring these skills from face-to-face interactions to a video-based link. Simply put, the video interactions were markedly richer, subtler and easier than the telephone interactions.

One implication of this finding is that video should be most helpful in those situations when people's rich set of interaction skills are most in demand. Our data suggest that one such case is the resolution of conflicts. Cultural norms tend to discourage people from handling disagreements directly, requiring people to rely more heavily on subtle unspoken cues to interpret another person's attitude. Through video, a speaker may notice an addressee's unconscious expression or shift in posture, and adjust her utterance in mid-stream to head off a misinterpretation. This finding suggests that, relative to audio only, video would also be of use for handling other highly interactive situations when nonverbal cues are most helpful, such as negotiating or creating rapport. Finally, video should be more effective than the phone for people who are working together from different locations over a long period of time. If remote collaborators can communicate richer information more easily, they are likely to have fewer misunderstandings and more effective interactions. Of course, it would be better still to carry out such activities face-to-face, but these are at least a few areas where video and audio offer an advantage over audio alone.

It is important to note that although these subtle cues arrive through the visual channel, participants often use the audio channel to respond to the information. For example, after *seeing* someone show doubt, the participants in our study often *verbally* explained more fully, asked about the other person's concern, etc. Consequently, this visual backchannelling from the listeners to the speaker might be thwarted by voice-activated video conferencing technologies, which switch everyone's video image to show the current speaker. It is most effective to enable the speaker to view the other participants as the others view the speaker.

Notice, also, that much of the speaker's adaptation depends on tightly integrated verbal exchanges. Previous studies have shown that small delays in the audio can seriously disrupt participants' ability to reach mutual understanding and reduce their satisfaction with the conversation [13,23]. This presents a design trade-off, because synchronizing video with audio is typically accomplished by delaying the audio until the more computationally-intensive video is processed. However, delaying the audio reduces the participants' ability to make use of the information in the video. In effect, delaying audio to provide synchronized video and audio generates a rich set of visual information, but people cannot effectively respond to it because of the introduced delay. We have found that users of such a system feel far more frustration about this delay than they do over of a lack of synchronization [23].

In our DVC prototype, we transmitted the audio as fast as possible, without attempting to preserve synchrony with the video. One-way audio delays ranged from .32 to .44 seconds, while video arrived noticeably later. We found that, although the participants found it slightly disturbing when the video did not match the audio, they still had well-timed interactions that were far richer than those we have observed among people using a commercial video conferencing system, which delayed the audio about .57 seconds (one-way) to synchronize with video [23]. In fact, one group who was using this audio-delayed commercial system decided to turn off the audio and use a half-duplex speaker phone connection instead, dramatically demonstrating their preference for instantaneous audio over synchronized audio and video.

It was somewhat surprising that the participants accomplished rich interactions using the DVC prototype with audio delays as long as .44 seconds. Still, our experience is consistent with a previous study that showed minimal detrimental effects of 0.3 second audio delays

(one way) compared to 0.9 second delays [13]. On the other hand, Wolf [27] found that participants who interacted with a .420 second one-way audio delay rated the audio and interaction quality significantly lower than those who experienced .167 second delays. However, that study reported only participants' *ratings* of audio quality and simultaneous speech rather than measuring *actual* audio problems and overlapping speech. Our experience concurs with Wolf's findings because our participants did notice and complain about the 0.32-0.44 second audio delays they experienced. Nonetheless, we found that they were able to effectively compensate for audio delays within that range.

LIMITATIONS OF VIDEO

Despite the many advantages of having a video and audio channel rather than just audio, a comparison of the desktop video conferences with face-to-face interactions revealed aspects of interactions that could not be accomplished through our DVC prototype, and in some cases, video in general. Interacting remotely through video makes it difficult or impossible for participants to: manage turn-taking, control the floor through body position and eye gaze, notice motion through peripheral vision, have side conversations, point at things in each other's space or manipulate real-world objects. Of course, some of these limitations may be overcome by providing additional capabilities, and we discuss these possibilities as design implications. On the other hand, some of these same drawbacks also create specific advantages. In particular, video interactions may not require as much social protocol and, in the case of desktop video conferencing, people can spontaneously draw upon resources in their own environments as the conversation unfolds.

Managing Turn Taking

The participant's turn-taking patterns were significantly different during face-to-face and DVC meetings. Specifically, in the five-way interactions, the participants exchanged more turns per minute when talking face-to-face than they did in DVC conversations ($F(1,314)= 43.28$, $p<.0001$), and their turns were shorter in duration ($F(1,250) = 7.13$, $p < .008$). In the two-way meetings, the participants again exchanged more turns per minute when face-to-face than when in a DVC ($F(1,76)= 5.14$, $p < .026$), but there was no difference in the duration of the turns. Table 2 shows the mean number of turns per minute and the mean duration for each condition.

Turns/min. / *Duration*	DVC	Face-to-face
Two-person conversation	**6.6** / *3.2*	**7.8** / *3.2*
Five-person conversation	**2.3** / *4.5*	**4.2** / *2.7*

TABLE 2. Average number of turns per minute and duration of turns in seconds during desktop video conferences (DVC) vs. face-to-face interactions in two- and five-person conversations.

Exchanging shorter turns more frequently indicates that in the face-to-face encounters, the participants were able to more tightly coordinate their utterances, which we know enhances their ability to reach mutual understanding [3,4,12]. It is unclear why the participants in the two-way DVC and face-to-face meetings did not differ in their turn duration even though they exchanged turns more rapidly. Apparently, there must have been more silence in between turns during the DVCs. Nonetheless, in both cases, the turn rate indicates that the participants more tightly coordinated their turn-taking. This finding indicates that while video improves the ability to handle conflict and confidential issues compared with the phone, face-to-face interactions are even better than video conferences for handling those types of sensitive issues.

It should be noted that this turn-taking finding is inconsistent with some other similar research. Sellen [19] did not find a significant difference in number of turns when comparing two video conditions to face-to-face interactions. It seems plausible that the difference stems from the fact that her video setup used analog audio and video over short distances, which resulted in nearly no transmission delay. This would suggest that difficulties in managing turn taking may be the result of an audio delay and not an inherent limitation of video. However, Krauss and Bricker [13] varied the audio transmission delay for an audio-only task, and they showed a difference in turn length only when the delay lasted .9 seconds, but not when there was no delay or a .3 second delay. He also found no difference in turn frequency in any condition. The apparent disagreement among these findings suggests the need for more research.

Controlling the Floor

In face-to-face interactions, we saw many instances when people used their eye gaze to indicate whom they were addressing and to suggest a next speaker [18]. In many instances when more than one person started speaking at the same time, the next speaker was determined by the eye gaze of the previous speaker. We even saw one interesting example of using a gesture to "reserve" a conversational turn. During a particularly active stretch of conversation, Jack and Jeff start speaking at the same time. As he speaks, Jeff reaches over and touches Kate's document to make a point about it. He loses the turn, but he keeps his finger on the document, essentially reserving his right to the next turn, which in fact he took. Others have also noted uses of gestures to *prevent* others from taking a turn [5].

In contrast, in our desktop video conferencing prototype, it was impossible to direct attention toward a specific person in a multiway conference. Everyone sees you through the same camera, so if you are looking at one person's video image, it appears to everyone as if you are looking at all of them. Not surprisingly, participants in DVCs did not seem to use body or eye position to control the floor. (But see Sellen [19] for one way to overcome this obstacle.)

Instead, people tended to use each other's names to address each other. For instance, at one point, Jack and Craig start talking at the same time and Jack gets the turn. As Jack starts speaking, Jeff overlaps with, "I didn't hear you, Craig." When Jack is done, Jeff again explicitly asks Craig to take the next turn. Had they been face-to-face, Jeff might have used gestures to help Craig win the previous turn from Jack.

Using Peripheral Cues

We observed many instances during face-to-face meetings when participants used their peripheral vision to notice a change in each other's body, head or eye position and then responded by coordinating their own activity. In DVCs, the video window on the screen is a small part of a participant's visual field. If a participant is not looking at or near that window, she is much less likely to notice motion in the window. Even large-scale motion on the other end, such as moving an arm to the face, translates into a small change in the remote participants' environment and can easily be missed if they are not looking near the video window. Changes in eye gaze are particularly unlikely to be noticed through peripheral vision.

For example, during a 30 second sequence of Jeff and Kate's face-to-face interaction, Jeff is talking and Kate is looking down as she takes notes. Three times, Jeff looks up at Kate for confirmation, and each time, she nods or replies "Yeah," without looking up or interrupting her writing. She is obviously able to sense his head position and eye gaze and recognize that he is seeking a response.

We did not see this kind of subtle coordination based on peripheral cues in DVCs. If anything, we saw many instances when the participants just missed each other's glances. (See Heath and Luff [11] for a discussion of similar problems.) In one typical example, Jeff glances at Kate as he finishes speaking, but looks away too soon to catch Kate's nod in response. At another point, Jeff misses Kate's smile, so he responds to her comment seriously.

Having Side Conversations

Side conversations were impossible using the desktop video conferencing prototype because people could not address particular participants and because everyone shared a single audio channel. The closest we observed was two participants using the channel to discuss topics of interest to themselves while the others waited for the conversation to become more general.

In the five-way face-to-face meeting, the conversation occasionally broke into two parallel conversations and then seamlessly transitioned back to a single conversation. For example, at one point Jack makes a joke and everyone but Kate laughs. While the others continue with the conversation, Kate looks at Jack and asks him to repeat what he said, which he does. She comments on his joke and then they both refocus on the group's conversation. This side conversation can be accomplished because participants can "open" a second audio channel and because the visual cues enable everyone to understand who is participating in which conversation when.

Pointing

If a participant in a DVC points to one of the video images on her screen, it is difficult for the others to use spatial position to figure out whom is being addressed. They can use only the verbal context to make an educated guess. Pointing can be used, however, to focus attention on certain parts of their own environments.

We saw few instances of pointing in either the two-way or five-way DVC, even to indicate items in their own space. We saw one instance when Jeff pointed to his image of the two people in Billerica, but from the other participants' perspective, he simply appears to be pointing to his screen. It is difficult for them to determine exactly which image he was indicating.

In contrast, we saw many instances of pointing during both face-to-face meetings. During the five-way meeting, the participants repeatedly pointed to places in their own documents and at times reached over to each other's documents to point out a particular line or diagram. In the two-way meeting, Kate pushes part of the document between her and Jeff, and they repeatedly point to different parts of it as they talk about it.

Manipulating Real-World Objects

The participants in our study never had the need to jointly observe, manipulate, or build an object, but these activities present such an obvious limitation to remote video conferencing that we point it out. However, during both the two-way and five-way face-to-face interactions, the participants did review hard copy documents. By observing their joint behavior with the documents, we noticed at least two limitations of video in this regard: it does not allow participants to build on each other's work, and it does not allow them to "look over each other's shoulders" to gain another perspective.

We saw instances of both of these during face-to-face interactions, whereas no equivalent behavior was possible in our DVCs. For example, when Kate pushes the document to the middle of the table, she and Jeff write and draw on it, at times building on each other's sketches or comments. They also continue to write on their own pads, transitioning easily between their own space and the shared space. In another simple example from the five-way meeting, Kate leans over to look at Jack's copy of the document to see where he is looking.

Advantages of Video over Face to Face

In addition to these limitations, we saw evidence of advantages of desktop video conferencing over face-to-face meetings. First, we found, as have others, that video conferencing distanced our participants because they could not make eye contact or use peripheral cues to pick up on subtleties [6,9,14]. As a result, there seems to be less of a pressure to carry out standard social practices that may make interactions "less efficient" [6]. When someone physically drops by, we are often expected to ask how they are and have an introductory social conversation before getting down to business. This type of interaction serves an important purpose, but it can be seen as reducing short-term efficiency. At least in those interactions when social chit-chat is less critical, people may choose to use a desktop video conference to help focus on the work at hand.

We see an interesting parallel with electronic mail, which people use, among other reasons, when they want to handle certain factual or practical matters, perhaps without "bothering" with accompanying social interaction. Using e-mail does not mean people do not also use other communication techniques to handle more social or interactional matters. It merely provides another option when textual content is most important.

Secondly, participants in DVCs are normally in their own offices, with all those resources at their disposal. All participants can spontaneously bring into the discussion both on-line and off-line materials if they become relevant. In addition, if one person is looking for something or handles an interruption (a phone call, a person dropping by or even an incoming e-mail message), the other members can draw on their own private space to use the time productively. As a result, meetings can and were used at times more like loose connections akin to sharing an office. In some cases, individual meetings smoothly shifted between focused conversations and loose, intermittent interactions. Users of other desktop video conferencing systems have also been reported to open up video connections between offices to create virtual shared offices, while at other times they used the connection for focused interactions [1,7].

This kind of interaction may be inappropriate at times, and in fact members of the team we observed said they were sometimes annoyed when one member stopped participating as he read or answered an incoming e-mail message. But this type of "shared space" can be a useful environment for certain types of activities.

Design Implications from the Disadvantages of Video

Comparing our desktop video conference system with face-to-face meetings highlighted the possible shortcomings of video for remote collaboration. In particular, participants found it difficult to manage turn-taking, control the floor, notice small movements through peripheral vision, have side conversations, point at things in each other's space and manipulate real-world objects. One approach to compensating for these limitations is to use electronic means to directly substitute for some of the interactional mechanisms

observed in face-to-face behavior. For example, one might provide an explicit visual mechanism for controlling the floor in group interactions or enable the ability to open a separate channel for side conversations.

One potential danger of such an approach is that it may force people to take explicit actions to carry out behaviors that are normally negotiated unconsciously. For example, requiring users to indicate explicitly when they want the next turn eliminates their ability to manage the politeness issues around floor control. Doing so may also eliminate cues about the degree of spontaneity and enthusiasm in a participants' desire to contribute. In addition, artificial behaviors may be interpreted differently by other participants. For instance, a person who would have been seen as enthusiastic might be perceived as dominating if she uses an explicit mechanism rather than a socially negotiated one to manage floor control.

In general, we recommend thinking in terms of enabling a new range of collaborative *tasks* by broadening the shared space among participants. Such a system may entail providing one or more mechanisms to enable particular collaborative activities (e.g. pointing, noticing motion), but it should also expand the participants' ability to handle collaboration issues through the standard social negotiation process.

We experimented with this approach by integrating a shared drawing program with the DVC. Previous studies had shown that the ability to draw shared diagrams and pictures is an important aspect of many interactions [16,24,25]. As part of the DVC prototype, we developed a program called Show Me, which allows users to share an image of anything they can display on their screens. They can draw on top of shared images, or construct a joint drawing from scratch. Within the shared drawing tool, users can type or draw at the same time, they can erase anyone else's work, and they can always see where everyone else is pointing with their cursors. (Show Me has since been modified and developed into a product [21].)

By increasing the nature of the shared space among the participants, Show Me enabled a wider range of collaborative activities not available through video. Not only could participants bring into discussion any document or image from their workstations, but they could also use the cursor to point to parts of the image, and they could track each other's attention through their cursors. We did not build in protocols to prevent people from erasing each other's work, relying instead on the audio connection and social negotiation for people to manage its usage. Our intention was to enable a new type of activity (shared drawing), which involved building technology to support certain behaviors (showing certain objects, pointing, tracking attention) as well as relying on existing collaborative behaviors to handle many of the social interaction issues.

On the other hand, the tool was not as successful as we would have liked because it allowed for sharing of only one bitmap image at a time. If two people wanted to jointly edit a document, they could not work on the actual document. One person would have to make changes and then transmit a bitmap of the updates. To move on to another page, one person had to page the actual document and then transmit the image of the next page. The essential problem was that the shared space was not as broad as we would have liked, and that limitation did appear to reduce the usefulness of the tool.

Our observations lead us to conclude that tools designed to supplement a video conferencing system should:

- enable behaviors associated with particular collaborative tasks;
- broaden users' shared environment;
- take advantage of users' existing collaboration skills;
- not require conscious actions for behaviors that are normally done unconsciously.

We should not try to use a video conferencing system to carry out tasks that require manipulating objects, pointing and other behaviors that are not fully supported through video alone. For example, it would be unwise to attempt to have a group video meeting about a controversial topic, expecting everyone to feel they had a chance to contribute. This situation depends too heavily on the ability to achieve smooth floor control among many people (and perhaps to have side conversations), which are weaknesses of a simple audio-video link. Similarly, it may be possible to use video to teach someone how to assemble a machine, but it will not be as effective as a face-to-face demonstration because both the participants could not point to and manipulate the objects together.

We hope that we have drawn attention to some of those limitations so that we may have more realistic expectations of video systems that do not specifically address them, and so that we may focus our development efforts on tools that help compensate for these drawbacks.

ACKNOWLEDGEMENTS

We acknowledge the Digital Integrated Media Environment and the Conferencing and Collaboration groups in Sun Microsystems Laboratories, Inc. for developing and studying the desktop video conferencing prototype described in this paper. We also thank Jonathan Grudin for his helpful comments on an earlier draft. We especially thank the members of the team we observed for their cooperation.

REFERENCES

1. Bly, S.A., Harrison, S.R., and Irwin, S., Media spaces: Bringing people together in a video, audio, and computing environment, In *Communications of the ACM*, 36(1), 1993, pp. 28-45.
2. Chapanis, A., Ochsman, R.B., Parrish, R.N. and Weeks. G.D. Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem-solving, *Human Factors*, 14 (6), 1972, pp. 487-509.
3. Clark, H.H. and Schaefer, E.F. Collaborating on contributions to conversations, *Language and Cognitive Processes*, 2(1), 1987, pp. 19-41.
4. Clark, H.H. and Wilkes-Gibbs, D. Referring as a collaborative process, *Cognition*, 22, 1986, pp. 1-39.
5. Duncan, S. Some signals and rules for taking speaking turns in conversation, *Journal of Personality and Social Psychology*, 23(2), 1972, pp. 283-292.
6. Fish, R.S., Kraut, R.E. and Chalfonte, B.L. The VideoWindows system in informal communications. In *Proceedings of the Conference on Computer-Supported Cooperative Work* (Los Angeles, CA, 1990), pp. 1-11.
7. Fish, R.S., Kraut, R.E. and Root, R.W. Evaluating Video as a technology for informal communication, *Proceedings of CHI '92 Human Factors in Computing Systems*, (Monterey, CA, 1982), pp. 37-48.
8. Francik, E., Ehrlich Rudman, S., Cooper, D. and Levine, S. Putting innovation to work: Adoption strategies for multimedia

communication systems, *Communications of the ACM*, 34(12), 1991, pp. 53-63.

9. Gale, S. Human aspects of interactive multimedia communication, *Interacting with Computers*, 2, 1990, pp. 175-189.

10. Grudin, J., Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces, *Proceedings of the Conference on Computer-Supported Cooperative Work*, (Portland, OR, 1988) pp. 85-93.

11. Heath, C. and Luff, P. Disembodied conduct: Communication through video in a multimedia environment. In *Proceedings of the CHI '91 Conference on Human Factors in Computing Systems* (Monterey, CA., 1992), pp. 99-103.

12. Isaacs, E. and Clark, H.H., References in conversation between experts and novices, *Journal of Experimental Psychology: General*, 116, 1987, pp. 26-37.

13. Krauss, R.M. and Bricker, P.D., Effects of transmission delay and access delay on the efficiency of verbal communication, *Journal of the Acoustic Society of America*, 41, 1967, pp. 286-292.

14. Mantei, M.M., Baecker, R.M., Sellen, A.J., Buxton, W.A.S. and Milligan, T., Experiences in the use of a media space, In *Proceedings of the CHI '91 Conference on Human Factors in Computing Systems* (New Orleans, LA, 1991), pp. 203-208.

15. Ochsman, R.B and Chapanis, A., The effects of 10 communication modes on the behavior of teams during co-operative problem-solving, *International Journal of Man[sic]-Machine Studies*, 6, 1974, pp. 579-619.

16. Olson, M.H. and Bly, S.A., The Portland experience: A report on a distributed research group, *International Journal of Man[sic]-Machine Systems*, 34(2), 1991, pp. 211-228. Reprinted: *Computer-supported Cooperative Work and Groupware*, Saul Greenberg (Ed.), London: Academic Press, pp. 81-98.

17. Pearl, A., System support for integrated desktop video conferencing, *Sun Microsystems Laboratories, Inc. Technical Report*, TR-92-4, 1992.

18. Sacks, H., Schegloff, E. and Jefferson, G., A Simplest systematics for the organization of turn-taking for conversation, *Language*, 50, 1974, pp. 696-735.

19. Sellen, A.J., Speech patterns in video-mediated conversations, *Proceedings of CHI '92 Human Factors in Computing Systems*, (Monterey, CA, 1982), pp. 49-59.

20. Short, J., Williams, E. and Christie, B., *The social psychology of telecommunications*, London: John Wiley & Sons, 1976.

21. *ShowMe(TM)*, a product of SunSolutions, a Sun Microsystems, Inc. Business, 1992. Free evaluation diskettes available by calling (800) 647-8333.

22. Tang, J.C., Involving social scientists in the design of new technology, *Taking Software Design Seriously: Practical Techniques for Human-Computer Interaction*, Karat, J. (Ed.), Boston: Academic Press, 1991, pp. 115-126.

23. Tang, J.C. and Isaacs, E.A., Why do users like video? Studies of multimedia-supported collaboration, *CSCW: An International Journal*, 1993 (in press). Also a *Sun Microsystems Laboratories, Inc. Technical Report*, TR-92-5, 1992.

24. Tang, J.C. and Minneman, S.L., VideoDraw: A video interface for collaborative drawing, *ACM Transactions on Information Systems*, 9(2), 1991, pp. 170-184.

25. Tang, J.C., Findings from observational studies of collaborative work, *International Journal of Man [sic]-Machine Studies*, 34(2), 1991, pp. 143-160. Reprinted: *Computer-supported Cooperative Work and Groupware*, Saul Greenberg (Ed.), London: Academic Press, pp. 11-28.

26. Williams, E., Experimental comparisons of face-to-face and mediated communication: A Review, *Psychological Bulletin*, 84(5), 1977, pp. 963-976.

27. Wolf, C.G., Video conferencing: Delay and transmission considerations, in *Teleconferencing and Electronic Communications: Applications, Technologies and Human Factors*, L.A. Parker and C.H. Olgren (Eds.), 1982.

28. Wulfman, C.E., Isaacs, E.A., Webber, B.L., and Fagan, L.M. Integration discontinuity: Interfacing users and systems, *Proceedings of Architectures for Intelligent Interfaces: Elements and Prototypes* (Monterey, CA, 1988) pp. 57-68.