Multimedia Systems (1994) 2:63–73

Multimedia Systems
© Springer-Verlag 1994

What video can and cannot do for collaboration: a case study

Ellen A. Isaacs*, John C. Tang**

SunSoft, Inc., 2550 Garcia Avenue, Mountain View, CA 94043, USA

Abstract. As multimedia become an integral part of collaborative systems, we must understand how to design such systems to support the user's rich set of existing interaction skills, rather than requiring people to adapt to arbitrary constraints of technology-driven designs. To understand how we can 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, the video channel of our desktop video conferencing prototype adds or improves the ability to show understanding, forecast responses, give nonverbal information, enhance verbal descriptions, manage pauses, and express attitudes. These findings suggest that video may be better than the phone for handling conflict and other interaction-intense activities. However, 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. Nonetheless, when compared with face-to-face interaction, 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 enable rich interactions fully, video should be integrated with other distributed tools that increase the extent and type of shared space in a way that enables natural collaborative behaviors within those environments.

Key words: Remote collaboration – Desktop video conferencing – Computer-supported cooperative work – User interfaces – Conversation

1 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 (Francik et al. 1991; Grudin 1988). When multimedia technology is included in collaborative systems, more design challenges are added, since so little is known about how to combine various media in ways that are effective and 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 (Wulfman et al. 1988).

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 video is often intuitively presumed to 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 (Chapanis et al. 1972; Gale 1990; Ochsman and Chapanis 1974; Short et al. 1976; Williams 1977).

Did these previous studies somehow miss finding the effect of video, or are our intuitions about the value of video misleading? By re-examining some of the assumptions and conditions of the previous studies, we identified three reasons why they might not have detected any effect of video in support of interaction.

Firstly, the previous studies measured the *product* (e.g., decisions, quality of solutions, completion times) of short, problem-solving interactions. The effects of video are more likely to be visible when studying the *process* of interactions. 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 (Clark and Schaefer 1987; Clark and Wilkes-Gibbs 1986; Isaacs and Clark 1987; Sacks et al. 1974; Williams 1977). If video is effective at enhancing the process of interaction, people may perceive their interactions to be more satisfying, and it may encourage coworkers to collaborate more frequently. 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.

Secondly, the effects of video in supporting interaction may be most visible over the long term, and may be too subtle or difficult to capture in short, laboratory experiments. As Gale (1990) notes:

* e-mail: isaacs@sun.com
** e-mail: tang@sun.com
Correspondence to: E.A. Isaacs

> 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.

We would expect that richer interactions would lead to more productive and/or higher quality results in the long term, although more research would be needed to test this hypothesis.

Thirdly, most of the previous studies were among strangers who were asked to accomplish an artificial task for the purposes of the study. That is, the participants did not have working relationships with each other, and were not dealing with an issue that highly motivated them. Yet the interactional cues that video transmits are likely to play a more important role among people who know each other and are accomplishing real work that requires complex social negotiation. Again, Gale (1990) notes:

> The results from this study suggest that by adding audio and video to the communications medium we allow groups to perform more 'social' activities. ... A possible reason for the lack of difference in the quality of the output of the groups is that the tasks used in this study were not sensitive to social factors.

Thus, previous studies may have missed the effects of video because they are too subtle to see among strangers carrying out impersonal tasks.

In this context, we believe there is still good reason to pursue video as an integral part of collaborative technology. To study the user-interface implications of using video for remote collaboration, we observed a team of engineers who were using a desktop video conferencing (DVC) prototype. The prototype enabled digital audio-video connections between workstation desktops. 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 potential benefit of video in remote conversations compared with audio alone, and to point out how video interactions can 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.

2 Method

We observed a team of five software engineers who were distributed across three sites. Two worked in a building in Billerica, Mass., two worked in a building in Mountain View, Calif., and one worked in another Mountain View building about 500 yds (ca. 450 m) 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.

The DVC prototype provided a simple interface for requesting a desktop conference that followed a telephone model for making connections. The interface enabled a user to request a conference with one or two other people, and those people had to accept that conference request before audio-video connections were made. The interface allowed users to specify what kind of connections they wanted for a conference: audio, video, and/or a shared drawing tool called Show Me. Show Me allowed users to share an image of anything they could display on their screens. They could draw on top of shared images or construct a joint drawing from scratch. Within Show Me, users could type or draw at the same time, they could erase anyone else's work, and they could always see where everyone else was pointing with their cursors.

The default setting was to request audio, video, and Show Me connections with one other participant. Once a conference request was accepted, several windows popped up on each participant's workstation screen. Each person saw a video window showing the image of the other participant and a smaller preview window of the video image being sent to the other person. A full duplex audio connection was made with the other person. A Show Me window also popped up on each screen. (In a three-way connection, each person would see an additional video window of the second remote participant.)

The DVC prototype ran on Sun computer workstations with a prototype add-on board that enabled real-time video capture, compression, and display. For audio, the prototype used the 8 KHz μlaw encoding that is built into the workstation. 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 × 128 pixels, although that resolution could be scaled to any arbitrary size window. The default video frame rate was 5 frames/s; 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). The quality of the video image was less than that of broadcast television and certain analogue desktop video conferencing systems such as Cruiser (Root 1988) (due to the lower refresh rate and lower effective resolution). But it was of higher quality than video conference systems than run over ISDN or conventional phone lines. More details on the technical description of the prototype can be found in Pearl (1992).

The data from this study was drawn from a larger study that observed the team's work activity under three conditions: (1) before installing the DVC prototype (to understand their baseline collaborative activity); (2) with the DVC prototype fully installed, including audio, video, and the shared drawing tool; and (3) with the video channel subtracted from the DVC prototype, leaving audio and the shared drawing capabilities still in place. [See Tang and Isaacs, (1993) for a further description of the results from the larger study.] Since the team that was observed was not involved in the development of the DVC prototype, they had never experienced using the prototype before it was installed during the second condition of the study. Each team member was given a short (less than 15 min) demonstration of the prototype to familiarize them with the prototype's capabilities and user interface.

Although we took many measures of their work activity, the data for this paper are based on videotapes of six inter-

Table 1. People involved in each observed interaction

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

actions in three modalities: two desktop video conferences, two face- to-face interactions and two telephone conferences. Comparing interactions among the same people using various tools enabled us to isolate the effects of the tools on their interactions better. One of the DVC meetings included all five group members (call them Kate, Jeff, Jack, Dave, and Craig) and one was between just Kate and Jeff. Likewise, 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.

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

The videotapes of the six interactions were analyzed in the tradition of interaction analysis (Tang 1991a; Tatar 1989) to look for any changes in pattern among the three modalities. The qualitative analysis involved creating a detailed account of all the interactional behaviors among the group members, with a particular emphasis on behaviors that took advantage of audio and visual cues. The quantitative analysis involved comparing the mechanics of conversational turn taking (i.e., duration of turn, frequency of turn changes) between the face-to-face and DVC modalities.

3 Benefits of video over audio only

An analysis of the videotapes brought out the benefit of video conferencing compared to audio only. 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.

3.1 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 (Clark and Schaefer 1987; Clark and Wilkes-Gibbs 1986; Isaacs and Clark 1987). Furthermore, they expect various degrees of feedback depending on the complexity of the topic (Isaacs and Clark 1987). Head nods are a subtle and nonintrusive way of conveying understanding (Duncan 1972), and they were used extensively throughout the DVCs. Participants nodded their heads to varying degrees and at varying rates, showing various 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 explained a technical issue. At first, Jeff tilted his head and looked puzzled, but eventually he gave a slight head nod as he grasped the concept. Then he sighed and shook his head, acknowledging the issue as difficult. All these subtle reactions gave Kate a running commentary on the state of Jeff's understanding. Later, Kate asked him to confirm his understanding of an idea and he said "Uh huh," but then he looked down and pursed his lips as he considered the issue. Kate proceeded 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 said, "...we should probably take, like, the first part of the meeting and just go through and see what questions you guys have." After a 3-s pause, he said, "Okay? Then you can at least get your questions answered" (1 s pause). "And then we can hit you up for stuff that we want to know" (1 s). "Okay?" (1 s) "All right?" Finally, Jeff said "Yep" and continued. With no visual feedback, Dave had to explicitly 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. Addressees could give visual feedback on the level of their understanding without interrupting the speaker. Without the video, the participants had to work harder to get much less information about each other's understanding.

3.3 Forecasting responses

In the DVC, the participants not only indicated their level of understanding, they also occasionally forecast 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 made a point and Jeff tipped his head left and right in a gesture indicating "sort of." When she finished, he started his turn with "Yeah, but..."

Later, Kate started to nod in response to Jeff's comment but then stopped abruptly, indicating she thought she agreed but now was not sure. When she gave no indication of agreement at the end of his utterance, he prompted her with "Right?" He seemed 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 following examples.

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 midcourse. Of course, addressees may recognize that their reactions are not being forecast and therefore explicitly express their reactions verbally. But doing so requires more effort, and thus people may be more inclined to let subtle problems pass. In particular, participants may prefer not to express disagreement verbally 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.

3.4 Enhancing verbal descriptions with gestures

We also observed a variety of cases in which DVC participants made nonarbitrary gestures that emphasized their points. For example, Kate made a succession of gestures during her conference with Jeff. She said, "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, as shown in Fig. 1). There's so many functions now, the diagrams get all (flicked 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.

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 (as when on the phone), they may not express verbally the subtleties conveyed through their inadvertent gestures.

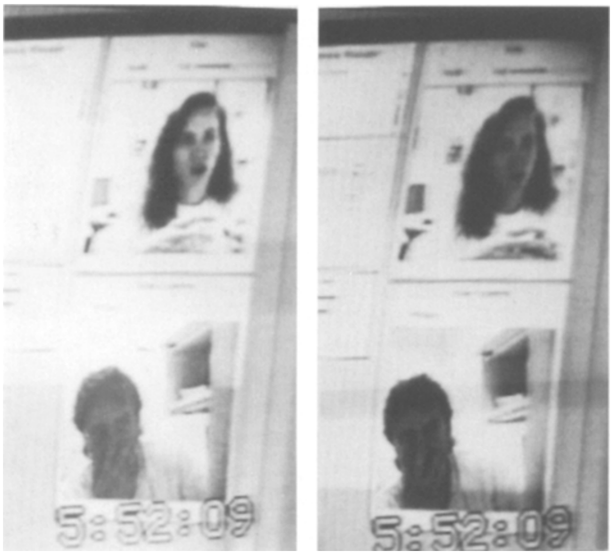

Fig. 1. Gesturing accompanying talking: a sequence of two images in time shows Kate (upper window) making a gesture to indicate *fit together*

3.5 Conveying purely nonverbal 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 meaningful gesture. For example, in the five-way DVC, Jack was frustrated about a decision, and asked "What does that benefit (this project)?" He then made a "zero" gesture with his hand and without saying any more. In the two-way DVC, Kate and Jeff finished discussing a problem that they were not in a position to resolve themselves. They looked at each other and made facial expressions that expressed "Oh well." Jeff shrugged and raised his hands, again as if to say, "such is life." They then moved 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 noticed that another person, Ted, was walking behind Craig and Dave as they were discussing a technical matter. Ted happened to be knowledgeable about the matter, so Jeff suggested asking him to join the conversation, which he did. 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.

The participants could not convey information purely nonverbally over the phone. One interesting incident occurred in the four-way phone conference. During this call, two of the participants, Jack and Kate, were in the same office sharing a speaker phone, and so they could see each other. At one point, Jeff asked Jack, "I forget, how big of a pain is it to add new built-ins, Jack?" After a 3-s pause Kate observed, "He doesn't look too happy," and Dave burst out with a laugh. Had Kate not been able to see Jack, the pause would have indicated only that he was considering the answer; Jack's spontaneous unhappy expression would not have been communicated.

3.6 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 gave a treatise on an issue as he leaned forward, moved his torso around and gestured with his arms. There could be no question about the strength of his conviction.

Later in this conference, Jeff told the group that he had written a software utility they could use. They expressed interest, but then Craig teased Jeff, "As usual, no documentation." Jeff

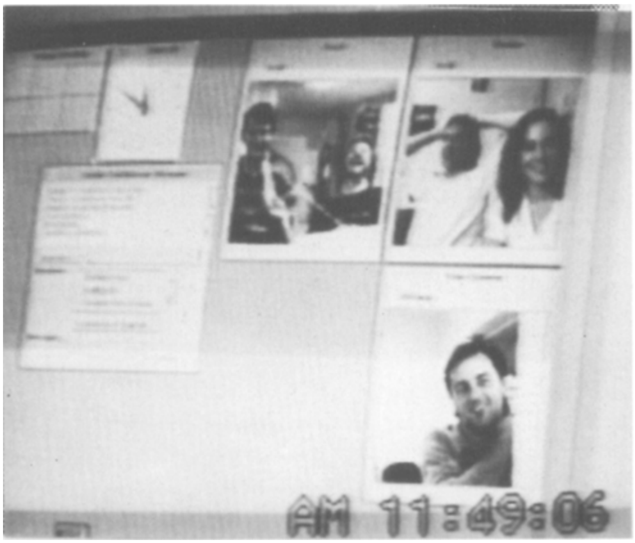

Fig. 2. Visually demonstrating humor: Craig (upper left video window) throws head back when others smile, showing appreciation of humorous response

smiled and said, "It's not even done yet!" Craig threw his head back while smiling broadly, as shown in Fig. 2. Jeff's words could be construed as defensive, but the smiles and Craig's response made it clear to everyone that the conversation was in fun. In contrast, in the three-way phone conversation, Jeff teased Craig about how his wife would react when she learned he was planning to spend 3 straight weeks on business trips. He got a minimal response that did not acknowledge the tease, so he exaggerated it. Again, he got a noncommittal response, so he explicitly asked Craig to address the tease. Craig did, but without revealing whether he found the whole topic amusing. Had Jim been able to see Craig, he would have been better able to interpret the response and adjust if, in fact, he had hit a sore spot.

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 that person was saying, sometimes returning their gaze as soon as they agreed with the speaker 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. Previous research shows that people prefer to use unspoken cues to handle topics that raise politeness issues because it enables them to handle potentially threatening situations more gracefully and effectively (Brown and Levinson 1978; Isaacs and Clark; 1990). Over the phone, either the disagreement or misunderstanding is never communicated or the addressee must raise it explicitly, which can make it clumsy to resolve.

Of course, losing the shades of meaning conveyed in expressions and body position does not often cause dramatic effects, although there are cases when it would be critical to realize someone is not agreeing or that a comment is intended to be humorous. Our basic argument is that in the course of a conversation, or series of conversations, seeing each other's facial expressions, gestures, and posture generally increases the participants' level of mutual understanding without requiring extra effort. Although any single instance of conveying these cues is expendable, their aggregation over time makes a substantial difference in the rapport and kind of interactions that occur between colleagues.

3.7 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 elaborating further. However, we observed instances where the video indicated other meanings for a pause. For example, in the two-way conference, Kate responded to a question by looking to her left and consulting her notes for 13 s. Meanwhile, Jeff waited without trying to clarify his question. At another point, Jeff agreed to do something, and then scribbled a note to himself for the next 12 s. Kate looked up, saw what he was doing and waited until he had finished.

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 spent more than 2 min looking for an electronic mail message while the others waited. There were extremely long pauses, punctuated by the other three teasing the two in Billerica and having a casual conversation among themselves. The Mountain View participants were able to monitor the other two members' progress and adjust their expectations accordingly.

There were certainly instances of nonproblematic pauses during the phone conference as well. In fact, one lasted as long as 28 s. However, on the whole they were more likely to be explained explicitly. At one point in a phone conversation, Dave said "I'm trying to look down things that are open bugs," meaning that he was consulting a list. For the next 7 min, his participation in the conversation was minimal, until he said "I can't find anything else in here."

3.8 Design implications of adding video

Our results clearly show that even with its mediocre quality, the low-bandwidth video used in our prototype provided a great deal of information that participants used to enhance their interactions relative to phone conversations. People have extensive experience interpreting small changes in expressions, gestures, and body position and adapting their responses. 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, relative to the phone, video should be most helpful in situations in which a 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 them to rely more on subtle unspoken cues to interpret person's attitude. Through video, speakers may notice addressees' unconscious expressions or shifts in posture, and adjust their utterances in midstream to head off misinterpretations. 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. Notice, also, that much of the speaker's adaptation depends on tightly integrated verbal exchanges. Previous studies show that small delays in the audio can seriously disrupt the participants' ability to reach mutual understanding and reduce their satisfaction with the conversation (Krauss and Bricker 1967; Tang and Isaacs 1993). 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 delay. We have found that users of such a system feel far more frustration about this delay than they do about a lack of synchronization (Tang and Isaacs 1993).

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 0.32 to 0.44 s, 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 by about 0.57 s (one-way) to synchronize with video (Tang and Isaacs 1993). 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, demonstrating their strong 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 0.44 s. Still, our experience is consistent with a previous study that showed minimal detrimental effects of 0.3-s audio delays (one way) compared to 0.9-s delays (Krauss and Bricker 1967). We note that Wolf (1982) found that participants who interacted with a 0.420-s one-way audio delay rated the audio and interaction quality significantly lower than those who experienced 0.167-s delays. However, that study reported only the 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-s audio delays. Nonetheless, we found that they were able to compensate effectively for audio delays within that range.

4 Limitations of video

Despite the many advantages of having a video-and-audio channel rather than just audio, a comparison of the DVCs 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. However, 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 DVC, people can spontaneously draw upon resources in their own environments as the conversation unfolds.

4.1 Managing turn taking

The participant's turn-taking patterns during face-to-face and DVC meetings were significantly different. In the five-way interactions, an analysis of variance showed that the participants exchanged more turns/min when talking face to face (4.2) than they did in DVC conversations (2.3) [$F(1{,}314) = 43.28$, $P < 0.0001$], and their turns were shorter in duration (2.7 s/turn 4.5 s/turn) [$F(1{,}250) = 7.13$, $P < 0.008$]. In the two-way meetings, the participants again exchanged more turns/min face to face (7.8) than in a DVC (6.6) [$F(1{,}76) = 5.14$, $P < 0.026$], but there was no difference in the duration of the turns (2.3 s/turn vs. 2.3 s/turn). Figure 3 graphs the mean number of turns/min and the mean duration for each condition. In this analysis, a turn length was defined as when a person started speaking to either when the speaker finished (e.g., denoted by a pause) or when someone else started speaking.

Exchanging shorter turns more frequently indicates that in the face-to-face encounters, the participants were able to coordinate their utterances more tightly, which, research has shown, enhances their ability to reach mutual understanding (Clark and Schaefer 1987; Clark and Wilkes-Gibbs 1986; Isaacs and Clark 1987). 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 in the face-to-face meetings

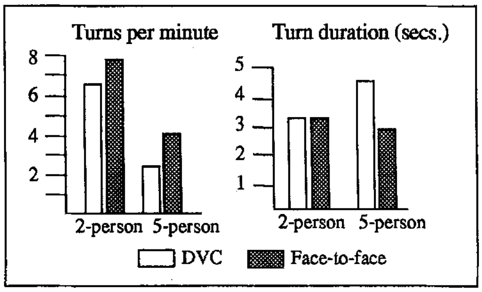

Fig. 3. Average number of turns/min and duration of turns/s during desktop video conferences (DVCs) versus face-to-face interactions in two- and five-person conversations

they exchanged turns more rapidly. Apparently, there was more silence between turns during the DVCs. Nonetheless, in both cases, the turn rate indicates that the participants coordinated their turn-taking more tightly. 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 similar research. Sellen (1992) 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 are primarily a result of the audio delay and not an inherent limitation of video. However, Krauss and Bricker (1967) varied the audio transmission delay for an audio-only task, and they showed a difference in turn length only when the delay lasted 0.9 s, but not when there was no delay or a 0.3-s delay. They also found no difference in turn frequency in any condition. The difference in our findings may also be caused by our different measurement of "turns." Sellen (1992) did not count as turns short "backchannels" that lasted less than 1.5 s, whereas we used a cruder definition of turns that included any utterance that lasted at least 0.5 s. Perhaps the participants in our study used more backchannels when they were face-to-face than when they were talking over the DVC, which could account for the discrepancy in our findings.

4.2 Controlling the floor

In face-to-face interactions, we saw many instances of people using their eye gaze to indicate whom they were addressing and to suggest a next speaker (Sacks et al. 1974). 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 started speaking at the same time. As he spoke, Jeff reached over and touched Kate's document to make a point about it. He lost the turn, but he kept his finger on the document, essentially reserving his right to the next turn, which in fact he took. Others have also noted the use of gestures to *prevent* others from taking a turn (Duncan 1972).

In contrast, in our desktop DVC prototype, it was impossible to direct attention toward a specific person in a multiway conference. Everyone sees a speaker through the same camera, so if the others are looking at the speakers video image, it appears to them that the speaker is looking at all of them. Not surprisingly, the participants of the DVCs did not seem to use body or eye position to control the floor. [However, see (Sellen 1992) for one way to overcome this obstacle.]

Instead, people tended to use names to address each other. For instance, at one point, Jack and Craig started talking at the same time and Jack got the turn. As Jack started speaking, Jeff overlapped with, "I didn't hear you, Craig" in an attempt to direct the next turn toward Craig. However, Jack held on to the turn, and after he finished speaking Jeff again explicitly asked 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.

4.3 Using peripheral cues

We observed many instances during face-to-face meetings in which the 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 our DVC, the video window on the screen was a small part of a participant's visual field. A participant who was not looking at or near that window was 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, translated into a small change in the remote participants' field of view and could easily be missed if that person was not looking near the video window. Changes in eye gaze were particularly unlikely to be noticed through peripheral vision.

For example, during a 30 s sequence of Jeff and Kate's face-to-face interaction, Jeff was talking and Kate was looking down as she took notes. Three times, Jeff looked up at Kate for confirmation, and each time, she nodded or replied "Yeah," without looking up or interrupting her writing. She was obviously able to sense his head position and eye gaze and recognize that he was 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 1991) for a discussion of similar problems.] In one typical example, Jeff glanced at Kate as he finished speaking, but looked away too soon to catch Kate's nod in response. At another point, Jeff missed Kate's smile, so he responded to her comment seriously.

4.4 Having side conversations

Side conversations were impossible with the DVC 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 flowed back to a single conversation. For example, at one point Jack made a joke and everyone but Kate laughed. While the others continued with the conversation, Kate looked at Jack and asked him to repeat what he said, which he did. She commented on his joke and then they both refocused on the group's conversation. This side conversation was accomplished because the participants could "open" a second audio channel and because the visual cues enabled everyone to understand who was participating in which conversation when.

4.5 Pointing

If a participant in our DVC pointed to one of the video images on the screen, it was difficult for the others to use spatial position to figure out who was being addressed. They could use only the verbal context to make an educated guess. Pointing could 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 appeared to be pointing to his screen. It was 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 pushed part of the document between her and Jeff, and they repeatedly pointed to various parts of it as they talked about it. We did see instances of this kind of pointing when the parit-icipants used the Show Me shared drawing tool, but it was, of course, accomplished through a cursor on the shared window rather than over the video channel.

4.6 Manipulating real-world objects

The participants in our study never needed to observe, manipulate, or build an object jointly, 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: (1) it does not allow participants to build on each other's work, and (2) 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 using our DVC, again unless they used the shared drawing tool. For example, when Kate pushed the document to the middle of the table, she and Jeff wrote and drew on it, at times building on each other's sketches or comments. They also continued to write on their own pads, moving easily between their own space and the shared space. In another simple example from the five-way meeting, Kate leaned over to look at Jack's copy of the document to see what he was looking at.

4.7 Advantages of video over face-to-face meetings

In addition to these limitations, we saw evidence of advantages of DVC 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 (Fish et al. 1990; Gale 1990; Mantei et al. 1991). As a result, there seems to be less pressure to carry out standard social practices that may make interactions "less efficient" (Fish et al. 1990). 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 (Whittaker et. al. 1994). Kraut et. al. (1990) reported that 20% of the face-to-face office conversations they observed consisted of "social, non-task-oriented conversation," whereas only 5% of video conference conversations were social. This type of social 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 DVC to help focus on the work at hand.

We see an interesting parallel with electronic mail, which people use, when, among other reasons, 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.

Participants in DVCs are normally in their own offices, with many resources at their disposal. All participants can spontaneously bring into the discussion both online and offline 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 DVC 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 (Bly et al. 1992; Fish et al. 1992).

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.

4.8 Design implications from the limitations of video

Comparing our DVC 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 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 that person uses an explicit mechanism rather than a socially negotiated one to manage floor control.

In general, we recommend thinking in terms of enabling a wider range of collaborative tasks by broadening the shared space among participants. This can be done by integrating other collaborative tools with a video-based system. 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.

The integration of the Show Me shared drawing program into the DVC prototype is an example of such an approach. Previous studies have shown that the ability to draw shared diagrams and pictures is an important aspect of many interactions (Olson and Bly 1991; Tang and Minneman 1991; Tang 1991b). We have mentioned some ways in which Show Me enabled a wider range of collaborative activities not available through video. It did so by increasing the nature of the shared space among the participants. Not only could participants bring any document or image from their workstations into discussion, 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.

However, the tool was not as successful as we would have liked because it allowed the sharing of only one bitmap image at a time. If two people wanted to edit a document jointly, 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:

- Broaden users' shared environment
- Enable behaviors associated with particular collaborative tasks
- 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 if they are not supported in the design of the system. For example, it would be unwise to attempt to use a simple video conference system 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 unless both the participants can point to and manipulate the objects together.

We hope that we have drawn attention to some 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. In addition, our study identified some specific ways in which video affects the *processes* rather than the *products* of interaction compared to audio only and face-to-face. It remains to be shown if and how these *process* effects accumulate over time into *productivity* or *product quality* effects. More research is needed to explore these longitudinal effects of video support for remote collaboration.

Acknowledgements. We acknowledge the Digital Integrated Media Environment and the Conferencing and Collaboration Groups in Sun Microsystems Laboratories, Inc. for developing and studying the DVC 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

Bly SA, Harrison SR, Irwin S (1993) Media spaces: bringing people together in a video, audio, and computing environment. Commun ACM 36:28–45

Brown P, Levinson S (1978) Universals in language usage: politeness phenomena. In: Goody E (ed) Questions and politeness, Cambridge University Press, Cambrdige, pp 56–311 University Press, 1978.

Chapanis A, Ochsman RB, Parrish RN, Weeks GD (1972) Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem-solving. Human Factors 14:487–509

Clark HH, Wilkes-Gibbs D (1986) Referring as a collaborative process. Cognition 22:1–39

Clark HH, Schaefer EF (1987) Collaborating on contributions to conversations. Language and cognitive processes 2:19–41

Duncan S (1972) Some signals and rules for taking speaking turns in conversation. J Personality Soc Psychol 23:283–292

Fish RS, Kraut RE, Chalfonte BL (1990) The videowindows system in informal communications. Proceedings of the Conference on Computer-Supported Cooperative Work, Los Angeles, Calif., pp 1–11

Fish RS, Kraut RE, Root RW (1992) Evaluating video as a technology for informal communication. Proceedings of CHI '92 Human Factors in Computing Systems, Monterey, Calif., pp 37–48

Francik E, Ehrlich Rudman S, Cooper D, Levine S (1991) Putting innovation to work: adoption strategies for multimedia communication systems. Commun ACM 34:53–63

Gale S (1990) Human aspects of interactive multimedia communication. Interacting with Computers 2:175–189

Grudin J (1988) Why CSCW applications fail: problems pn the design and evaluation of organizational interfaces. Proceedings of the Conference on Computer-Supported Cooperative Work, Portland, Ore., pp 85–93

Heath C, Luff P (1991) Disembodied conduct: communication through video in a multimedia environment. Proceedings of the CHI '91 Conference on Human Factors in Computing Systems, New Orleans, La., pp 99–103

Isaacs EA, Clark HH (1987) References in conversation between experts and novices. Journal of Experimental Psychology: General 116:26–37

Isaacs EA, Clark HH (1990) Ostensible invitations. Language in society 19:493–509

Krauss RM, Bricker PD (1967) Effects of transmission delay and access delay on the efficiency of verbal communication. J Acoustic Soc Am 41:286–292

Kraut RE, Fish RS, Root RW, Chalfonte BL (1990), Informal communication in organizations: form, function and technology. In: Oshkamp S, Spacapan S (eds) People's Reactions to Technology. Sage Publications, Newbury Park, pp 145–199

Mantei MM, Baecker RM, Sellen, AJ, Buxton, WAS, Milligan T (1991) Experiences in the use of a media space. Proceedings of the CHI '91 Conference on Human Factors in Computing Systems, New Orleans, La., pp 203–208

Ochsman RB, Chapanis A (1974), The effects of 10 communication modes on the behavior of teams during co-operative problem-solving. Int J Man-Machine Studies, 6:579–619

Olson MH, Bly SA (1991) The Portland experience: a report on a distributed research group. Int J Man-Machine Systems 34:211–228. Reprinted in: Greenberg S (ed) Computer-supported Cooperative Work and Groupware, Academic Press, London, pp 81–98

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

Root RW (1988) Design of a multimedia vehicle for social browsing. Proceedings of the Conference on Computer-Supported Cooperative Work, Portland, Ore., pp 25–38

Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50:696–735

Sellen AJ (1992) Speech patterns in video-mediated conversations. Proceedings of CHI '92 Human Factors in Computing Systems, Monterey, Calif., pp 49–59

Short J, Williams E, Christie B (1976) The social psychology of telecommunications. John Wiley, London

Tang JC (1991a) Involving social scientists in the design of new technology. In: Karat J (ed) Taking software design seriously: practical techniques for human-computer interaction. Academic Press, Boston, pp 115–126

Tang JC (1991b) Findings from observational studies of collaborative work. Int J Man-Machine Studies 34:143–160. Reprinted in: Greenberg S (ed) Computer-supported Cooperative Work and Groupware, Academic Press, London, pp 11–28

Tang JC, Isaacs EA (1993) Why do users like video? Studies of multimedia-supported collaboration. CSCW: Int J 1:163–196

Tang JC, Minneman SL (1991) VideoDraw: a video interface for collaborative drawing, ACM Trans Inform Syst 9:170–184

Tatar D (1989) Using video-based observation to shape the design of a new technology. SIGCHI Bulletin 21:108–111

Whittaker S, Frohlich D, Daly-Jones O, Informal workplace communication: What is it like and how might we support it?, Proceedings of the CHI '94 Conference on Human Factors in Computing Systems, Boston, Mass., pp 131–137

Williams E (1977) Experimental comparisons of face-to-face and mediated communication: a review. Psychol Bulletin 84:963–976

Wolf CG (1982) Video conferencing: delay and transmission considerations. In: Parker LA, Olgren CH (eds) Teleconferencing and electronic communications: applications, technologies and human factors, University of Wisconsin Extension Center for Interactive Programs, Madison, Wisconsin, pp 184–188

Wulfman CE, Isaacs EA, Webber BL, Fagan LM (1988) Integration discontinuity:rfacingusersandsystems.ProceedingsofArchitectures for Intelligent Interfaces: Elements and Prototypes, Monterey, Calif., pp 57–68

Dr. Ellen Isaacs works at SunSoft in the Human Interface Engineering group. She spends part of her time working with the Collaborative Computing Group doing research on multimedia-supported collaborative work, and the rest of her time designing user interfaces for a variety of collaborative and single-user applications. Before coming to Sun, she worked at a Stanford University lab on a project to develop a speech interface to a medical expert system. She received her PhD in cognitive psychology from Stanford, where she studied language use and collaboration in conversation, and her bachelors from Brown Unversity, where she studied psychology and semiotics.

Dr. John Tang works in the Collaborative Computing Group, an advanced development group within the Human Interface Engineering department at SunSoft, Inc. He studies collaborative work activity in order to guide the design and development of multimedia collaborative systems. Prior to joining Sun, John worked at Xerox PARC developing and studying several shared drawing prototype tools. His doctoral research at Stanford University was on studying group design activity. He received his degrees from Stanford University in the Mechanical Engineering Department, Design Division.