The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration

Roel Vertegaal
Cognitive Ergonomics Department
Twente University
The Netherlands
roel@acm.org

ABSTRACT

In this paper, we discuss why, in designing multiparty mediated systems, we should focus first on providing non-verbal cues which are less redundantly coded in speech than those normally conveyed by video. We show how conveying one such cue, gaze direction, may solve two problems in multiparty mediated communication and collaboration: knowing who is talking to whom, and who is talking about what. As a candidate solution, we present the GAZE Groupware System, which combines support for gaze awareness in multiparty mediated communication and collaboration with small and linear bandwidth requirements. The system uses an advanced, desk-mounted eyetracker to metaphorically convey gaze awareness in a 3D virtual meeting room and within shared documents.

KEYWORDS: CSCW, multiparty videoconferencing, awareness, attention, gaze direction, eyetracking, VRML 2.

INTRODUCTION

With recent advances in network infrastructure and computing power, desktop video conferencing and groupware systems are rapidly evolving into technologically viable solutions for remote communication and collaboration. Video conferencing is no longer limited to expensive circuit-switched ISDN networks and is starting to be used over standard internet connections in conjunction with groupware software. The central premise for the use of video mediated communication over traditional telephony has been that video images improve the quality of communication between individuals by increasing the available sensory bandwidth. In a face-to-face situation, auditory, visual and haptic expressions are freely combined to convey messages and regulate interaction. It has been presumed that by adding video to an audio-only communication link, mediated communication would bear a significantly closer resemblance to face-to-face communication. Firstly, we will show why this is not necessarily so. Secondly, we will show how designing mediated systems is a problem of conveying the least redundant cues first. We will show that by providing the right cues, one problem emerging from usability studies into multiparty video mediated communication may be solved: the difficulty of establishing who is talking or listening to whom in multiparty communication. With regard to cooperative work, we extend this notion to the problem of knowing who is talking about what. The central issue here is that regardless of whether audio or video is used, in multiparty communication and collaboration one should provide simple (i.e., unobtrusive and low-bandwidth), yet effective means of capturing and metaphorically representing the attention participants have for one another and their work [21, 30]. We will demonstrate that gaze direction is a good way of providing such information and review candidate solutions. Finally, we present the eye-controlled GAZE Groupware System, a virtual meeting room which supplements multiparty audio conferencing with gaze awareness, allowing users to see where other participants look, be it at each other or within documents.

CHI '99 Pittsburgh PA USA

CONVEYING THE RIGHT CUES

Face-to-face communication is an extremely rich process in which people have the ability to convey an enormous amount of information to each other. In mediating the process of human communication, it is not obvious that such information richness is easily replicated by adding video images to standard telephony. Indeed, empirical studies [23] show the difference between face-to-face communication and video mediated communication to be significantly greater than the difference between video mediated communication and audio mediated communication. We may indeed attribute such findings to the large difference in sensory bandwidth between face-to-face and mediated conditions. Sensory bandwidth is characterized by the number of cues (actions which convey information from one human to another) conveyed by the different media. Verbal cues are the actual words spoken in a conversation, non-verbal cues include the way in which these words are spoken (paralinguistic speech), facial expressions, gaze, gestures, bodily movement, posture and contact, physical proximity and appearance [2]. Theoretically, the notion that we can simulate face-to-face situations under mediated conditions is a correct one. In practice, however, it seems that the number of cues that need to be conserved in order to accomplish a complete replication is far greater than one would expect. Simply adding video is only a minor step. And in conditions where much of the information is redundantly coded it seems to actually be an insignificant step where it comes to improving regulation of conversations or task performance [23]. The notion that the addition of video images should make mediated communication significantly more like face-to-face communication may have been based on a misinterpretation of Short et al.'s Social Presence Theory [24]. In this theory, communication media are ranked according to the degree in which participants feel co-located. Face-to-face communication would provide the greatest sense of social presence, followed by video, multispeaker audio and monaural audio. This ranking was based on a factor analysis of subjective ratings of dyadic (two-person) conversations using the various media, and does indeed suggest that the amount of social presence is improved by increasing the number of cues conveyed. So why then does the addition of video images to audio-only communication seem to be an insignificant step towards replicating face-to-face conditions where it comes to regulation of conversations or task performance? We believe

this may, to a large extent, be attributed to a typical redundant coding scheme for those visual cues that are conveyed by a single stream of video. As Short et al. themselves pointed out, when cues are redundantly coded, we can no longer predict the effects of a communication system upon interaction by listing differences in the number of cues conveyed by different media. For example, a speaker preparing to yield the floor to a listener may use a combination of the following expressions: completion of a grammatical clause; a sociocentric expression such as 'you know'; a drawl on the final syllable; a shift in pitch at the end of the phonemic clause; a drop in loudness; termination of a hand gesture; relaxation of body position; and resumption of eyegaze towards the listener [9, 15, 24]. Note that we see a merging of verbal, paralinguistic, gestural, postural, and gaze-related cues, all indicating the same thing. When confronted with a different medium, speakers may easily adapt their behaviour by using different combinations of cues or by simply dropping several cues without failing to yield the floor. Indeed, half of the non-verbal cues in the above example are auditory, and five of the total of eight cues could be conveyed by telephone. This makes it extremely hard to find differences between video mediated communication and audio mediated communication in terms of performance in a joint task or, for that matter, more objective variables of conversational structure such as number of interruptions, duration of simultaneous speech or number of utterances. Indeed, empirical studies have so far failed to find clear differences in terms of conversational structure or task performance between video- and audio mediated communication (for an excellent overview, see [23]). When improving mediated communication, should we therefore aim to model face-to-face conditions even closer? We agree with Dennett [7] that it is not very realistic to think that face-to-face situations can, or indeed should be substituted by modeling the world on a one-to-one basis (a question already raised by Descartes [8]). We conclude that we should avoid putting too much research emphasis on improving mediated communication by means of increasing the bandwidth for video, and first focus on providing non-verbal cues which are less redundantly coded in speech, thereby hoping to provide some essential characteristics of face-to-face communication without intending to substitute it completely.

PROBLEMS WITH MEDIATING MULTIPARTY COMMUNICATION

In multiparty conditions (in which more than two persons communicate), gaze direction may well serve as a good example of such a cue. Multiparty conversational structure is much more complicated than its dyadic equivalent. As soon as a third speaker is introduced, the next turn is no longer guaranteed to be the non-speaker. When the number of participants rises beyond three, it becomes possible to have side conversations between subgroups of people. This can pose problems for the regulation of, for example, turntaking. When we consider the above example of a speaker yielding the floor in a multiparty situation, the question arises to whom he would like to yield the floor. With the notable exception of gaze direction (or rather the general orientation of body, head and eyes) and perhaps pointing gestures, such attention-related information is not coded by the eight cues listed in the above example. It can only be conveyed by telephone by means of explicit verbal references (e.g., calling someone by name) or the internal context of conversation. We believe turntaking problems with current multiparty conferencing systems (regardless of whether they use video or audio) may be attributed to a lack of cues about other participants' attention.

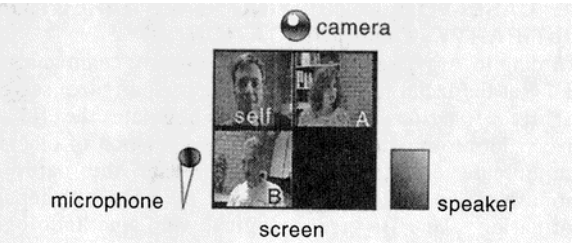

Figure 1. A single-camera video conferencing system.

Isaacs and Tang [13] performed a usability study of a group of five participants using a typical desktop video conferencing system. They found that during video conferencing, people needed to address each other by using each others names and started to explicitly control the turntaking process by requesting individuals to take the next turn. In face-to-face interaction, however, they saw many instances when people used their eyegaze to indicate whom they were addressing and to suggest a next speaker. Often, when more than one person started speaking at the same time, the next speaker was determined by the eyegaze of the previous speaker without the need for conventions or explicit verbal intervention. Similarly, O'Connaill et al. [19] found that in video conferencing more formal techniques were used to achieve speaker switching than in face-to-face interaction. They too attribute this to the absence of certain speaker-switching cues. This suggests that multiparty communication using video conferencing is not necessarily easier to manage than using telephony. Single-camera video systems such as the one shown in figure 1 do not convey deictic visual references to objects (such as the computer screen) or persons (such as the other participants) outside the frame of reference of the camera any more than telephony. To some extent, the participants' presumption that video conferencing is more like face-to-face interaction than telephony may actually lead to inappropriate use of such visual cues. Isaacs and Tang [13] show how, when a participant 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. Similarly, subjects may try to establish eye contact by gazing at the video image of a participant. Although the large angle between the camera and the screen usually prevents looking each other in the eyes (as one would need to look at the camera and the video image simultaneously), even if they were to establish eye contact, they would establish it with every participant in the group.

With respect to the conservation of gaze direction, we therefore identified the following incremental requirements for conferencing systems [29, 31]:

1) Relative position: Relative viewpoints of the participants should be based on a common reference point (e.g., around a shared workspace), providing basic support for deictic references. One may add a corresponding spatial separation of audio sources (e.g., by means of stereo panning) in order to ease selective listening, for example during side conversations.

2) Head orientation: Conveying the general orientation of looking may help participants in achieving deixis (e.g., "What do you think?'), and may provide basic support for knowing who is attending to whom.

3) Gaze (at the facial region): Conveying the exact position of looking within each other's facial region may also help in achieving deixis, and may provide support for knowing whether others are still attending. Mutual gaze constitutes eye contact.

THE CASE FOR CONVEYING GAZE DIRECTION IN MULTIPARTY COMMUNICATION

According to Argyle and Kendon, in two-party communication, looking at the facial region of the other person (gaze) serves at least five functions [2, 3, 15]: to regulate the flow of conversation; to provide feedback on the reaction of others; to communicate emotions; to communicate the nature of relationships; and to avoid distraction by restricting input of information. Due to technological and methodological complications, most studies into the role of gaze direction in communication were limited to two-person (dyadic) situations. In the early seventies, Argyle [2] estimated that when two people are talking, about 60 percent of conversation involves gaze, and 30 percent involves mutual gaze (or eye contact). People look nearly twice as much while listening (75%) as while speaking (41%). The amount of gaze is also subject to individual differences such as personality factors and cultural differences. For example, an extrovert may gaze more than an introvert. Also, there is more gaze in some kinds of conversations than others. If the topic is difficult, people look less in order to avoid distraction. If there are other things to look at, interactors look at each other less, especially if there are objects present which are relevant to the conversation [4]. In general, however, gaze is closely linked with speech. According to Kendon [15], person A tends to look away as she begins a long utterance, and starts looking more and more at her interlocutor B as the end of her utterance approaches. This pattern should be explained from two points of view. From the first point of view, in looking away at the beginning, person A may be withdrawing her attention from person B in order to concentrate on what she is going to say. When she approaches the end of her utterance, the subsequent action will depend largely upon how person B is behaving, necessitating person A to seek information about her interlocutor. From the second point of view, these changes in gaze can come to function as signals to person B. In looking away at the beginning, person A signals that she is about to begin an utterance, forestalling any response from person B. Similarly, in looking at person B towards the end of her utterance, she may signal that she is now ceasing to talk yet still has attention for him, effectively offering the floor to person B.

So how do these results hold in a multiparty condition? We conducted a study into the synchronization between auditory/articulatory attention and gaze at the facial region in four-person conversations [31]. When someone is listening or speaking to an individual, there is indeed a high probability that the person she looks at is the person she listens (88% chance) or speaks to (77% chance). In this more or less dyadic condition we found percentages of gaze similar to those found by Argyle, with about 1.6 times more gaze while listening than while speaking. However, when a person starts speaking to all three listeners, she will typically distribute her gaze over all of them. In this condition we found that the total percentage of gaze (while speaking) rises to 59% of the time. We may therefore conclude that gaze is indeed an excellent cue for establishing who is talking or listening to whom in multiparty face-to-face communication. In the next section, we investigate whether a representation of this cue can improve multiparty mediated communication.

Empirical Evidence in Multiparty Mediated Conditions

Very few, if any, studies exist in which the isolated effect of representing gaze direction in multiparty mediated communication has been empirically evaluated. Sellen [22] examined the differences in conversational structure between three multiparty conditions: using face-to-face communication; using a single-camera desktop video conferencing system (similar to the one depicted in figure 1); and using a Hydra system: a setup with multiple cameras, monitors and speakers which preserves relative position (including separation of audio), head orientation and, according to Sellen, gaze (Hydra [23] will be discussed further on in this paper). Although Sellen found differences in terms of objective measures (such as amount of simultaneous speech) between face-to-face and mediated conditions, she did not detect any differences between the two mediated systems. Sellen attributed this, in part, to the small screens of Hydra and their separation. As Heath and Luff [12] pointed out, movements in the periphery of vision which appear on a screen lose their power to attract attention. In addition, the still-present angle between camera and monitor in Hydra, albeit small, may have inhibited correct perception of gaze at the facial region [31]. Qualitative data did indicate subjects preferred the Hydra system over single-camera video conferencing. Reasons given included the fact that they could selectively attend to people, and could tell when people were attending to them. They also confirmed that keeping track of the conversation was the most difficult in the single-camera video conferencing condition. However, such conclusions may, in part, also be attributed to the separation of audio sources in the Hydra system.

We investigated the isolated effect of representing gaze directional cues on multiparty mediated communication, relative to the availability of other nonverbal upper-torso visual cues (see [31] and future publications). Groups of three participants (2 actors and 1 subject) solved language puzzles under three mediated conditions (all of which conveyed audio):

1) Motion video only, showing actor gaze 14% of time.
2) Motion video with head orientation, showing actor gaze 7% of time.
3) Head orientation and appearance only, showing actor gaze 32% of time. Actors manually selected one of three still images for display: looking at subject; looking at other actor, and looking at a computer terminal.

We found no effect of gaze directional cues, or any other nonverbal upper-torso visual cues, on task performance. However, gaze directional cues in the form of head orientation caused the number of deictic verbal references to persons used by the subjects to increase significantly by a factor two. We found a significant increase in turn frequency of about 25% in condition 3. A significant positive linear relationship between the amount of actor gaze at the facial region of subjects and the number of subject turns ($r=.34$, $p<.02$) and speaker switches ($r=.37$, $p<.01$) accounted for this finding. Thus, representing the gaze direction of the actors increased turn frequency, but only if it could be recognized by subjects as being aimed at themselves (i.e., as gaze at their facial region). We found no effect on turntaking of other nonverbal upper-torso visual cues. These findings suggest that all our requirements for multiparty conferencing systems should be met (i.e., they should preserve relative position, head orientation and gaze).

We conclude that although there have not been enough studies into the isolated effect of gaze directional cues on mediated multiparty communication, our evidence suggests that conveying gaze direction — especially gaze at the facial region — eases turntaking, allowing more speaker turns and more effective use of deictic verbal references. Depending on the task situation, however, this does not necessarily result in a significant performance increase. In the next section, we will investigate the role of gaze directional cues in collaboration.

THE CASE FOR CONVEYING GAZE DIRECTION IN COOPERATIVE WORK

We have so far examined the role of gaze direction in multiparty communication. Although some studies have investigated the role of looking at things during face-to-face collaboration, there are, to our knowledge, few empirical studies examining the effect of conveying gaze direction during computer supported cooperative work. Argyle and Graham [4] found that if a pair of subjects were asked to plan a European holiday and there was a map of Europe in between them, the amount of gaze dropped from 77 percent to 6.4 percent. 82 percent of the time was spent looking at the map. Even when they presented a vary vague, outline map, subjects looked at it for 70% of the time, suggesting that they were keeping in touch by looking at and pointing to the same object, instead of looking at each other. They also found there was little attention for the map if it was irrelevant to the topic of conversation.

Within the realm of computer supported cooperative work, Ishii and Kobayashi [14] demonstrated how the preservation of relative position and the transfer of gaze direction could aid cooperative problem solving through their ClearBoard system. They conducted an experiment in which two participants were asked to solve the "river crossing problem", a puzzle in which two groups of people (typically missionaries and cannibals) should reach the other side of a river with certain restrictions on who can join whom in the boat. According to the authors, the success of this game depends heavily on the point-of-view of the players. Participants could see video images of each other through a shared drawing board on which they could also sketch the problem. Ishii and Kobayashi concluded that it was easy for one participant to say on which side of the river the other participant was gazing and that this information was useful in jointly solving the problem. Colston and Schiano [6] describe how observers rated the difficulty people had in solving problems, based upon their estimates of how long a person looked at a particular problem, and how his or her gaze would linger after being told to move on to the next problem. They found a linear relationship between gaze duration and rated difficulty, with lingering as a significant factor. This suggests that people may use gaze-related cues as a means of obtaining information about the cognitive activities of a collaborator. Velichkovsky [27] investigated the use of eyetracking for representing the point of gaze during computer supported cooperative problem solving. Two people were asked to solve a puzzle represented on their screen as a random combination of pieces which had to be rearranged using the mouse. The two participants shared the same visual environment, but the knowledge about the situation and ability to change it on the way to a solution were distributed between them. One of the partners (the expert) knew the solution in detail but could not rearrange the pieces. The other (the novice) could act and had to achieve the goal of solving the puzzle without having seen more than a glance of the solution. In the first condition, they could only communicate verbally. In the second condition, the gaze position of the expert was added by projection into the working space on the screen of the novice. In the third condition, the expert used his mouse instead to show the novice the relevant parts of the task configuration. Both ways of conveying the attention of the partners improved performance. The absolute gain in the case of gaze position transfer was about 40%. Approximately the same gain was obtained with mouse pointing. In a second experiment, the direction of gaze position transfer was reversed from the novice to the expert. Here too, a significant gain was found in the efficiency of distributed problem solving. Apparently, experts could see the types of barriers novices confront in their activity and were therefore able to give more appropriate advice. This shows that gaze position transfer may be useful in situations where manual deixis is impossible: the novices could not use their mouse for pointing because they needed it to manipulate puzzle pieces.

We conclude that although the effect of providing a representation of gaze direction in cooperative work may be highly dependent on the task situation, a closer coordination between the communication and cooperation media with respect to conserving such deictic cues can be considered beneficial. In the next section, we will review existing systems in which gaze directional cues are preserved.

PREVIOUS SOLUTIONS

Over the years, a number of multiparty conferencing systems have been developed which complied with the earlier presented design recommendations. Such systems preserved relative position (including spatial separation of audio), head orientation and gaze. Negroponte [18] describes a system commissioned by ARPA in the mid-1970s to allow the electronic transmission of the fullest possible sense of human presence for five particular people at five different sites. Each of these five persons had to believe that the other four were physically present. This extraordinary requirement was driven by the government's emergency procedures in the event of a nuclear attack: the highest ranking members of government should not be hiding in the same nuclear bunker. His solution was to replicate each person's head four times, with a life-size translucent mask in the exact shape of that person's face. Each mask was mounted on gimbals with two degrees of freedom, so the 'head' could nod and turn. High-quality video was projected inside of these heads. In this rather humorous setup, each site was composed of one real person and four plastic heads sitting around a table in the same order. Each person's head position and video image would be captured and replicated remotely. According to Negroponte, this resulted in lifelike emulation so vivid that one admiral told him he got nightmares from these 'talking heads'. A technical advantage of this system was that only one camera was needed at each site to capture the video image of the participant's head, resulting into only one stream of video data from each participant (we will further address this issue below). It may, however, be difficult with this system to capture gaze at the facial region correctly. A technical disadvantage was the elaborate setup of the talking heads: the total number of heads required is almost the square of the number of participants ($n^2 - n$; in which n is the number of participants).

Sellen [23] describes the Hydra system, a setup of multiple camera/monitor/speaker units in which relative position (including spatial separation of audio), head orientation and gaze might be preserved during multiparty videoconferencing. Hydra simulates a four-way round-table meeting by placing a box containing a camera, a small monitor and speaker in the place that would otherwise be held by each remote participant. Each person is therefore presented with his own view of each remote participant, with the remote participant's voice emanating from his distinct location in space. This way, when person A turns to look at person B, B is able to see A turn to look towards B's camera. According to Sellen, eye-contact (i.e., mutual gaze) should be supported because the angle between the camera and the monitor in each unit is relatively small. The separation of audio in the Hydra system may ease selective listening, allowing participants to attend to different speakers who may be speaking simultaneously. Although Hydra is of course a very elegant alternative to Negroponte's

system, it has some disadvantages. One disadvantage seems to be that the system does not preserve gaze at the facial region accurately enough [23]. Another disadvantage is that although participants can see when someone is looking at a (shared) workspace, their estimation of where this person looks within that workspace would probably be worse than possible with, e.g., Negroponte's system. A more technical drawback is that each camera in the setup provides a unique video stream, and that the number of cameras required is almost the square of the number of participants ($n^2 - n$; in which n is the number of participants). For three participants, only six Hydra units are needed, but when this number rises to five, twenty Hydra units are required. In a Multicast network [10], the bandwidth requirements of traditional single-camera video conferencing systems are greatly reduced. With Multicasting, a video stream of an individual user is not sent to individual remote participants by means of multiple connections. Instead, that video stream is 'broadcast' to all other participants simultaneously, requiring only one unit of the total network bandwidth at any time. With the Hydra system, such compression cannot be achieved, causing the amount of network bandwidth used to convey video to rise with almost the square of the number of participants ($n^2 - n$). This may have an effect on usability, as it may lead to problems with proper conveyance and synchronization of audiovisual information.

Okada et al.'s MAJIC system uses a rather more elaborate setup in an attempt to achieve a seamless integration of life-size images of the other participants with each participant's real work environment [20]. In essence, it is a bigger version of the Hydra system, with a more precise positioning of cameras, behind the monitors (i.e., a big video tunnel [1]). In each office, a thin half-transparent curved projection screen is placed behind a computer terminal in front of the user. On this screen, life-size video images of the other participants are projected. Behind each projection screen, video cameras are located at the center of the projected facial region of the other participants, one camera for each participant. This way, head orientational information is conveyed, and users may achieve eye-contact by looking at each other's faces, as long as those faces do not move too much relative to the camera lens behind them [1, 31]. A corresponding placement of microphones and speakers is used to ease selective listening. We may well consider the MAJIC system the closest we will get to replicating a face-to-face situation without holographic projection. However, the disadvantages of MAJIC are similar to those of the Hydra system. In addition, due to the large image size, each video stream will require considerably more bandwidth than with the Hydra system, assuming resolution is maintained.

A more recent development has been the embodiment of chat participants in virtual environments [5, 17]. Although such systems do include ways to pictorially represent users in a spatial formations, we will not elaborate upon them as they do not include a direct way of capturing gaze direction. For a discussion on the issues concerning such Collaborative Virtual Environments, we refer to [11].

We conclude that most systems which convey gaze direction in multiparty communication have difficulties preserving gaze at the facial region under all circumstances, and suffer from an inefficient use of network resources. They may also have difficulties preserving gaze awareness in the workspace. In the next section, we will describe our own candidate solution based on a direct measurement of gaze direction.

THE GAZE GROUPWARE SYSTEM

Based on our design recommendations we developed a prototype multiparty mediated communication and collaboration system which provides awareness about the participants' gaze position without some of the drawbacks of earlier systems. Instead of conveying gaze direction by means of multiple streams of video, the GAZE Groupware System (GGS) measures directly where each participant looks using an advanced desk-mounted eyetracking system [33]. The system represents this information metaphorically in a 3D virtual meeting room and within shared documents. The system does this using the Sony Community Place [26] plug-in, which allows interactive 3D scenes to be shared on a web page using a standard multiplatform browser such as Netscape. The GAZE Groupware System can be used in conjunction with any multiparty speech communication facility such as an internet-based audio conferencing tool, or standard telephony.

The Virtual Meeting Room

The GAZE Groupware System simulates a four-way round-table meeting by placing a 2D image (or *persona*) of each participant around a table in a virtual room, at a position that would otherwise be held by that remote participant. Using this technique, each person is presented with a unique view of each remote participant, and that view emanates from a distinct location in space. Each persona rotates around its own x and y axes in 3D space, according to where the corresponding participant looks. Figure 2 shows the system in use in a four-way situation. When Robert looks at Roel, Roel sees Robert's persona turn to face him. When Robert looks at Harro, Roel sees Robert's persona turn towards Harro. Based on our earlier findings, this should effectively convey whom each participant is listening or speaking to.

When a participant looks at the shared table, a lightspot is projected onto the surface of the table, in line with her persona's orientation. The color of this lightspot is identical to the color of her persona. This "miner's helmet" metaphor enables a participant to see exactly where the others are looking within the shared workspace. With their mouse, participants can put document icons on the table representing shared files. Whenever a participant looks at a document icon or within the associated file, her lightspot will be projected onto that document icon. This allows people to use deictic references for referring to documents (e.g., "Here, look at these notes"). Shared documents are opened by double clicking their icon on the table. When a document is opened, the associated file contents appears in a separate frame of the web page (see figure 2). In this frame, an editor associated with the file runs as an applet. When a participant looks within a file, all participants looking inside that file can see a lightspot with her color projected over the contents. This lightspot shows exactly what this person is reading. Again, this allows people to use deictic references for referring to objects within files (e.g., "I cannot figure this out"). We realize, that providing such information may invade the privacy of individual users. By (annoyingly) projecting their own gaze position whenever it is shared, we hope to ensure that individuals are aware their

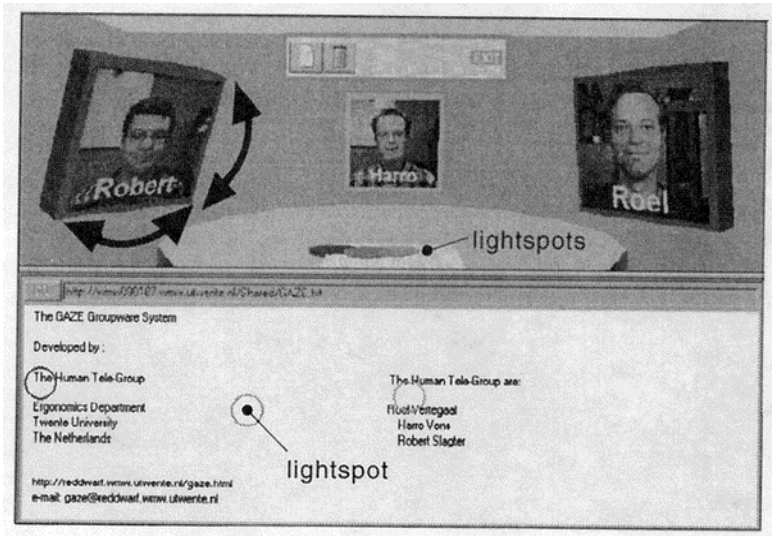

Figure 2. Personas rotate according to where users look.

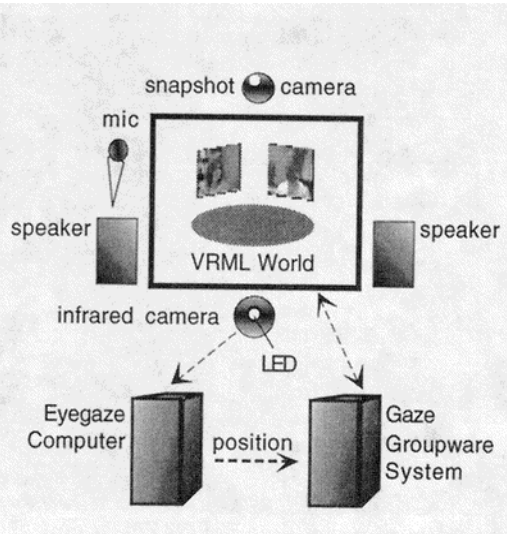

Figure 3. The GAZE hardware setup.

gaze position is transferred to others. Although files can be referred to by URL, they are currently restricted to ascii text.

HARDWARE SETUP

Each participant has a hardware setup similar to the one shown in figure 3. The GAZE Groupware System consists of two key components: the Eyegaze system, which determines where the participant looks; and the GGS computer, a Windows '95 Pentium running Netscape, the GAZE Groupware System, a web server, frame grabbing software and an internet-based audio conferencing tool. The Eyegaze system, which is discussed in detail below, reports the gaze position of the user over a serial link to the GGS computer. The GGS computer determines where the participant looks, manipulates her persona and lightspot, and conveys this information through a TCP/IP connection via a server to the other GGS computers. The Eyegaze system is not required. Participants can also switch to using their mouse to indicate point of interest. The video conferencing camera on top of the monitor is currently used to make snapshots for the personas (future versions will also incorporate motion video). When making a snapshot, it is important that users look into the video conferencing camera lens, as this will allow them to achieve a sense of eye contact during meetings.

THE LC TECHNOLOGIES EYEGAZE SYSTEM

When the eye remains relatively still for more than about 120 milliseconds, we speak of a fixation [28]. For determining where the user is looking, it is these fixation points that we are interested in. Our system measures the eye fixation points of a user by means of the Eyegaze System [16], an advanced, desk-mounted, imaging eyetracker with a spatial resolution of approximately 0.5 degrees of arc and a temporal resolution of 50-60 Hz. The Eyegaze system consists of a 486 computer processing the images of a high-resolution infrared video camera. This camera unit is mounted underneath the screen of the user (see figure 3), and is aimed at one of his eyes (see figure 4). On top of the camera lens, an infrared light source is mounted which projects invisible light into the eye. This infrared light is reflected by the retina, causing a bright pupil effect (the large circle in figure 5) on the camera image. The light *is* also reflected by the cornea of the eye, causing a small glint to appear on the camera image (the small dot in figure 5). Because the cornea is approximately spherical, when the eye moves, the cornea1 reflection remains roughly at the same position. However, the bright pupil moves with the eye. By processing the image on the computer unit, the vector between the center of the pupil and the cornea1 reflection can be determined. In order to correctly translate this vector into screen coordinates, the user needs to calibrate the Eyegaze system once before use. This calibration procedure takes about 15 seconds. When the coordinate remains within a specified range for approximately 120 msec (3 complete camera frames), the Eyegaze system decides that this is a fixation. It then starts reporting the coordinates over a serial link to the GAZE Groupware System running on a separate computer (see figure 3). The GAZE Groupware System uses this coordinate to determine at which object or participant on the screen the user is looking.

SOFTWARE IMPLEMENTATION

The GAZE Groupware System was implemented using the *Virtual Reality Modeling Language* 2.0 [25]. This cross-platform standard separates 3D graphic descriptions (rendered natively) from their dynamic behaviour (running on a JAVA Virtual Machine). Sony Community Place [26] is a plug-in for Netscape which implements the VRML 2 standard and adds a client-server architecture for sharing 3D graphics and behaviour over TCP/IP. For each dynamic object a user sees in the virtual meeting room, there is a corresponding JAVA object. Whenever such an object does something, its behaviour is broadcast via the Community Place Server by means of messages to the other systems. This way, all participants' copies of the meeting room are kept in sync. Eyetracker input is obtained from a small native driver application polling the serial port or the mouse. Document editors are JAVA applets running separately from the VRML world, although they do communicate with it to obtain eyetracking data and URLs. All code, graphics, and documents are shared using web servers running on each GGS computer.

EVALUATION OF THE SYSTEM

Informal sessions with several hundred novice users at ACM Expo'97 indicated our approach to be a promising one. Most participants seemed to easily interpret the underlying metaphors. The eyetracking technology was, in many cases, *completely* transparent. Users would sit behind the system and immediately start chatting, without calibration or instruction. As we spent most of our time empirically evaluating the underlying assumptions of the system (as discussed earlier, and in [31]), the prototype has not yet been tested for usability.

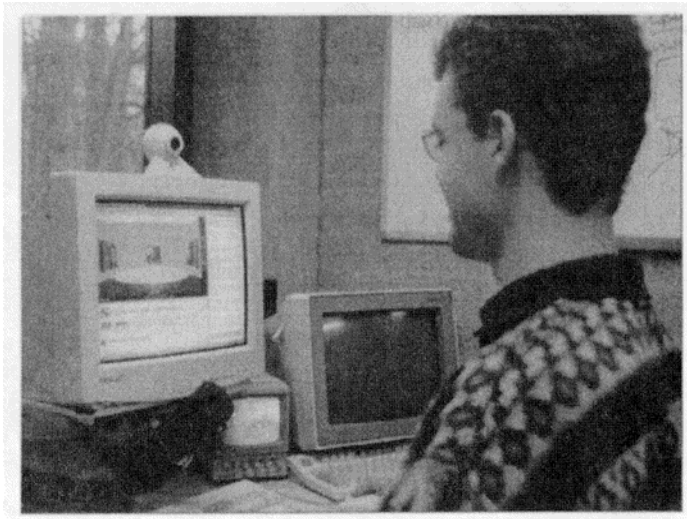

Figure 4. Participant using the GAZE Groupware System

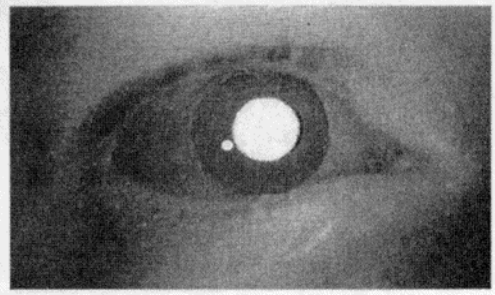

Figure 5. The infrared camera image.

Potential usability issues include:

- *No spatial separation of audio or visual encoding of speech activity.* Although it is not necessary for audio sources to be exactly co-located with the visual representation of users, spatial separation of their voices may ease selective listening [23]. This feature is not yet integrated into the prototype. Users currently need to depend on auditory discrimination of voices for identifying the source of individual speech activity. Spatial separation of audio and visual encoding of speech activity (using the animation techniques demonstrated in [32] or by using motion video) may solve this issue.
- *No option for motion video.* Although this is the subject of further investigation, the capturing of gaze at the facial region and the conveyance of motion video seem to be conflicting demands. This is because humans seem extremely sensitive to parallax [31]. Even when video tunnels are used, it is difficult to keep cameras and eye regions aligned at all times [1, 31]. Evidence presented in this paper suggests one should typically choose to convey gaze at the facial region. However, in future versions we hope to resolve this conflict, and convey both gaze and motion video in a network-scalable fashion.
- *Color coding and lightspot confusion.* Lightspots can only be attributed to a persona by color and synchronized movement. We would like to devise a more redundant coding scheme. When there are many lightspots, novices may get confused or distracted. It should at the very least be possible to turn lightspots off.
- *Privacy.* Although knowing what others are reading may be beneficial during a joint editing process, there are many task situations where this could be detrimental. Users should always be aware when their gaze is being transmitted, and when not. Currently, we hope to ensure this by (annoyingly) projecting the user's own gaze whenever she looks at shared objects. This is not a satisfactory solution.
- *Eyetracker limitations.* Although the eyetracker works well while talking, head motion is still limited to about 5-10 cm in each direction in order for gaze to be conveyed correctly (if the eye moves out of range, the eyetracker stops sending coordinates until it is back in range). However, similar restrictions apply to most other conferencing systems which convey gaze by means of video tunnels [1, 31]. In addition, a version of the Eyegaze System which allows 30 cm of head movement in each direction is almost ready for release. Although the eyetracker works fine with most glasses and contact lenses, a small percentage of users has problems with calibration. Eyetracking is still expensive, 'but current developments lead towards eyetrackers which are just another input device: inexpensive and transparent in use.
- *Meeting room restrictions.* Although this is not an intrinsic limitation, the system currently allows only four users in the meeting room. Users are currently not allowed to move freely through space or control their point of view, as this complicates the mapping of their gaze coordinates.

CONCLUSIONS

Designing multiparty mediated systems is a problem of conveying the least redundant cues first. Many of the cues conveyed by video are redundantly coded in speech. Less redundantly coded cues such as gaze direction are, however, usually not conveyed. We have shown how conveying gaze direction may solve a problem in multiparty mediated communication: establishing who is talking or listening to whom. Gaze direction is an excellent cue for providing such information. Evidence suggests that conveying gaze direction — especially gaze at the facial region — eases turntaking, allowing more speaker turns and more effective use of deictic verbal references. However, this does not necessarily result in a significant task performance increase. Depending on the task situation, gaze direction may also help establishing who is talking about what in cooperative work. Our GAZE Groupware System combines support for gaze awareness (preserving relative position, head orientation and gaze) in multiparty mediated communication and collaboration with small and linear bandwidth requirements. The system measures directly where each participant looks using a desk-mounted eyetracker. It represents this information metaphorically in a 3D virtual meeting room and within shared documents.

ACKNOWLEDGEMENTS

Thanks to Harro Vons, Robert Slagter, Gerrit van der Veer, Nancy and Dixon Cleveland, the LC Technologies team, Boris Velichkovsky, and Robert Rathbun of Cyberian Outpost for their important contributions to this project.

REFERENCES

1. Acker, S. and Levitt, S. Designing videoconference facilities for improved eye contact. *Journal of Broadcasting & Electronic Media* 31(2), 1987.
2. Argyle, M. *The Psychology of Interpersonal Behaviour.* London: Penguin Books, 1967.
3. Argyle, M. and Cook, M. *Gaze and Mutual Gaze.* London: Cambridge University Press, 1976.
4. Argyle, M. and Graham, J. The Central Europe Experiment - looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour* 1, 1977, pp. 6-16.
5. Benford, S., Greenhalgh, C., Bowers, J., Snowdon, S., and Fahlén, L. User Embodiment in Collaborative Virtual Environments. In *Proceedings of CHI'95.* Denver, Colorado: ACM, 1995.
6. Colston, H. and Schiano, D. Looking and Lingering as Conversational Cues in VMC. In *Proceedings of CHI'95.* Denver, Colorado: ACM, 1995.
7. Dennett, D. *Conciousness Explained.* London: Penguin, 1991.
8. Descartes, R. *Discours de la méthode.* Leiden: Jean Maire, 1637.
9. Duncan, S. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23, 1972.
10. Ericksson, H. MBONE: The Multicast Backbone. Communications of *ACM* 37(8), 1994, pp. 54-60.
11. Harrison, S. and Dourish, P. Re-Place-ing Space: The Roles of Place and Space in Collaborative Systems. In *Proceedings of CSCW'96.* Cambridge, MA.: ACM, 1996, pp. 67-76.
12. Heath, C. and Luff, P. Disembodied conduct: Communication through video in a multi-media office environment. In *Proceedings of CHI'91.* New Orleans: ACM, 1991.
13. Isaacs, E. and Tang, J. What video can and can't do for collaboration: a case study. In *Proceedings of Multimedia'93.* Anaheim, CA: ACM, 1993.
14. Ishii, H. and Kobayashi, M. ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. *Proceedings of CHI'92.* Monterey, CA: ACM, 1992.
15. Kendon, A. Some Function of Gaze Direction in Social Interaction. *Acta Psychologica* 32, 1967, pp. 1-25.
16. LC Technologies. *The Eyegaze Communication System.* Fairfax, VA, 1997. *http://www.lctinc.com*
17. Nakanishi, H., Yoshida, C., Nishimura, T., and Ishida, T. Freewalk: Supporting Casual Meetings in a Network. In *Proceedings of CSCW'96.* Cambridge, MA.: ACM, 1996, pp. 308-314.
18. Negroponte, N. *Being Digital.* New York: Vintage Books, 1995.
19. O'Connaill, B., Whittaker, S., and Wilbur, S. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human Computer Interaction* 8, 1993, pp. 389-428.
20. Okada, K., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In *Proceedings of CSCW'94.* Chapel Hill, NC.: ACM, 1994, pp. 385-393.
21. Raeithel, A. and Velichkovsky, B.M. Joint Attention and Co-Construction of Tasks: New Ways to Foster User-Designer Collaboration. In Nardi, B. (Ed.), *Context and Conciousness.* Cambridge, MA: MIT Press, 1996.
22. Sellen, A.J. Speech Patterns in Video-Mediated Conversations. In *Proceedings of CHI'92.* Monterey, CA: ACM, 1992, pp. 49-59.
23. Sellen, A.J. Remote conversations: the effects of mediating talk with technology. *Human Computer Interaction* 10(4), 1995.
24. Short, J., Williams, E., and Christie, B. *The Social Psychology of Telecommunications.* London: Wiley, 1976.
25. Silicon Graphics, Sony and ISO. The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding. *ISO/IEC 14772-1:1998,* 1998. *http://www.iso.ch/*
26. Sony. *Sony Community Place,* 1997. *http://vs.spiw.com/vs/*
27. Velichkovsky, B.M. Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition,* 3(2), 1995, pp. 199-222.
28. Velichkovsky, B.M., Sprenger, A., and Unema, P. Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem". In *Proceedings of INTERACT'97.* Sydney, Australia, 1997.
29. Vertegaal, R. Conversational Awareness in Multiparty VMC. In *Extended Abstracts of CHI'97.* Atlanta, GA: ACM, 1997, pp. 6-7.
30. Vertegaal, R., Velichkovsky, B.M., and Van der Veer, G. Catching the Eye: Management of Joint Attention in Cooperative Work. *SIGCHI Bulletin* 29(4), 1997.
31. Vertegaal, R. *Look Who's Talking to Whom.* PhD Thesis. Cognitive Ergonomics Department, Twente University, The Netherlands, 1998. ISBN 90 3651 1747.
32. Vertegaal, R. GAZE: Visual-Spatial Attention in Communication. Video Paper. In *Proceedings of CSCW'98.* Seattle, WA: ACM, 1998.
33. Vertegaal, R., Vons, H., and Slagter, R. Look Who's Talking: The GAZE Groupware System. In *Summary of CHI'98.* Los Angeles, CA: ACM, 1998.