In *Multimedia Tools and Applications,* Vol 1, March 1995.

The Illustrated Conversation

Judith S. Donath

MIT Media Lab

Abstract

Collaboration-at-a-Glance is a program which provides a visual interface to an online conversation. Although the participants in the conversation may be at widely separate locations, the interface provides a visible shared electronic space for their interactions. The participants each have a first person view-point from which they can see who else is present and who is communicating with whom. In the first part of this paper, I describe the current implementation. In the second part of the paper, I discuss some of issues involved in expanding it to supplement an ongoing conversation with additional expressive information.

1 Introduction: the electronic conversation

Many forms of collaboration and communication among groups of people online are essentially conversations[19][18]. These include the non-real time exchanges of postings found in netnews or on bulletin boards, and also the real-time discussions that take place in the social MUDs[1], on Internet Relay Chat [17] and in the "chatrooms" of America OnLine. These conversations are entirely text-based. Lacking visual cues, the participants in these conversations cannot see how many people are around, nor can they see where the attention of the group is directed. The participant cannot discern, at a glance, *who* else is present.

In the first part of this paper I will describe *Collaboration-at-a-Glance*, which provides a visual interface to an online conversation. Its focus is on visualizing presence and attention. In the second part, I will discuss some directions for future research, including issues such as the visualization of emotion and the balance of control between the viewer and the subject.

[1] A MUD (Multi-User Dungeon) is a "network-accessible, multi-participant, user-extensible virtual reality whose user interface is entirely textual"[7]. Some MUDs are used for game-playing. Many, however, are social environments where the primary active is real-time communication.

2 *Collaboration-at-a-Glance*

A scenario[2]**:** Lindsey is working at her computer – editing a screenplay, reading the news. Among the windows on her screen are several which show groups of faces, turned toward each other as if in conversation. Occasionally, there is movement in one of these windows – a head turns to face a different person. At one point, one of the window grows quite active. Many of the faces in it turn first towards one person, then to another; below the heads, a text window fills with messages. Lindsey is curious: this window is the marketing group with whom she works closely. The discussion, it turns out, is about a proposal to stage surprise bicycle stunts in shopping malls to promote their new feature – an idea she thinks is ludicrous. She clicks on the face of the idea's main proponent, Arthur, and types her objections.

On forty screens scattered across the country, in the window showing the marketing group, Lindsey's head turns to face Arthur's and the forty other participants in the argument read her remarks.

There is no picture of Arthur on Arthur's screen. Instead, he sees the picture of Lindsey looking straight out at him. He responds to her comments – and on all the screens, Arthur turns to face Lindsey (and on her screen, this means he looks directly out at her).

Meanwhile, the discussion continues, sometimes in public, sometimes privately. Martin, who is new to the company, asks for his friend John's opinion before he ventures a suggestion. Their conversation is a private aside within the group – they are seen conversing, but the contents of their notes are not included in the general text, appearing only on each other's screens.

Lindsey goes back to her editing. She's curious to hear what Susan, the producer of the film, will say about the proposed stunts. But Susan is not around. Her image appears as a stylized drawing, which means that she has her window set to record, and may review the discussion later, but is not actually present to participate.

2.1 Social visualization

The *Collaboration-at-a-Glance* window on Lindsey's screen provides her with a casual connection to her co-workers. Not only can she quickly see who is around, she can also see when and where an interesting conversation is taking place. The participants in the conversation know to whom they are talking. In contrast to many email and news-reading systems, those who are just listening are also visible.

The bandwidth requirements of *Collaboration-at-a-Glance* are extremely low: no images are sent across the network, only data about the state of the group. The pictures that the participants see are synthesized locally; they are a visualization of the data about the group's interactions. Yet *Collaboration-at-a-Glance* is not simply a low-bandwidth substitute for video conferencing. If limited bandwidth were not an issue – if one could have live video images of all of one's co-workers running simultaneously – *Collaboration-at-a-Glance* would not be redundant. *Collaboration-at-a-Glance* creates a simple movie of an unfilmable event: a meeting among widely separated people. The coherent 3D space

2. *Collaboration-at-a-Glance* is an implemented program. The scenario, however, is entirely fictional.

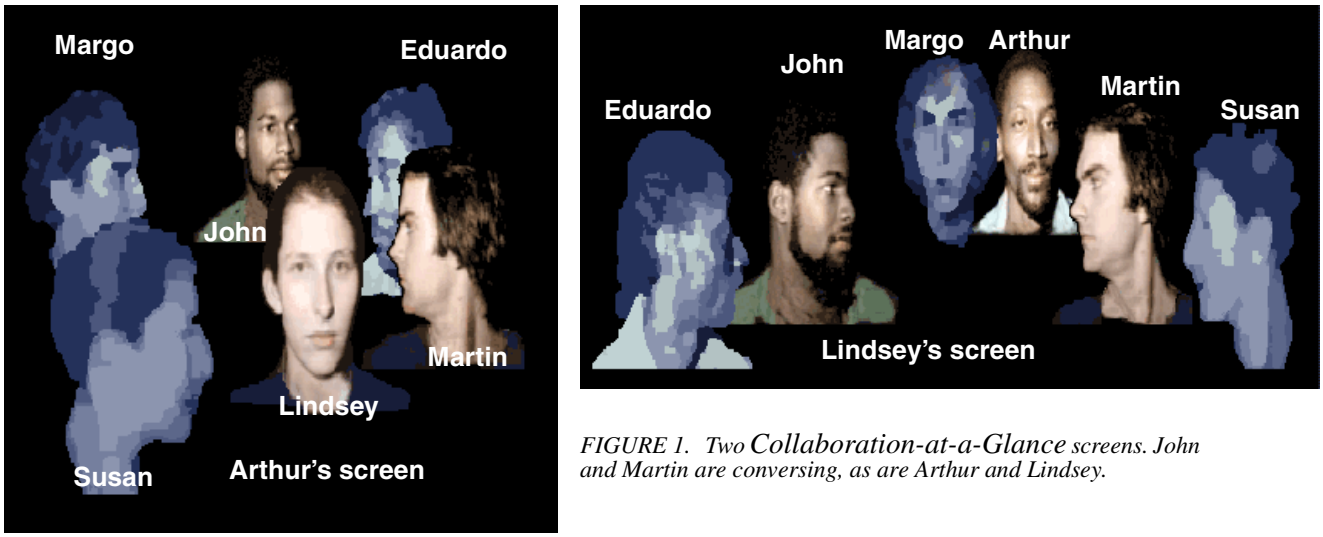

FIGURE 1. Two Collaboration-at-a-Glance screens. John and Martin are conversing, as are Arthur and Lindsey.

inhabited by these images shows the interactions between participants in a way that individual video windows cannot.

Representation	Meaning
Gaze direction	Attention & communication
Image style	Presence / absence
Location	Viewer preference
Image features	Physical appearance

Table 1: *Collaboration-at-a-Glance* representations

Collaboration-at-a-Glance maps abstract relationships and states of being to concrete visual representations. Some mappings are intuitively obvious, others are arbitrary; some show a range of values along a continuum, others the presence or absence of a particular quality. The choice of mappings is the fundamental issue in designing the visualization.

Attention. The mapping of gaze direction to attention is quite intuitive. The angle of clear vision in humans is quite small, subtending about *x* degrees of arc. In order to clearly see the subject of one's attention, one must be looking close to straight at it. Furthermore, turning to face the person or item of interest is an instinct that develops quite early. We are quite good at following gaze [12]. In general, we assume that someone's thoughts are on the item they are looking at.

It is not surprising that one is especially attuned to being looked at - we are able to detect with an accuracy of one minute of visual angle whether or not someone is looking at us [12]. Yet in many collaborative interfaces, the images of the other participants are always shown facing out at the viewer. In *Collaboration-at-a-Glance* gaze direction provides information. Usually, the people on the screen are facing in a variety of directions – everyone faces the viewer only if he or she actually is the center of attention.

FIGURE 2. (above) An image set of 7 frames, as used in the initial implementation - and in the color plates.

Presence. The mapping of image style - drawing or photograph - to degree of presence is open to a wider range of interpretations. In *Collaboration-at-a-Glance* photographic images are used to show a fully participating member of the conversation. These images appear much more life-like and immediate than stylized drawings: using the photographs to indicate an active user is not an arbitrary choice. The interpretation of the more abstract images, however, is application dependent. They can be used as placeholders for absent group members, or they can indicate passive listeners in an otherwise active group – auditors in a virtual classroom, audience members in an electronic panel discussion, software agents sent out to record online conversations.

Other kinds of processing are appropriate for providing other types of information. For instance, if we would like to make it clear who is very active in the conversation, versus those who may have been logged in for much of it, but who have been idle for quite some time, it would make sense to fade the image of those who have been idle for a long stretch. Fading can show a continuum; it also emphasizes the images of those who are active.

While the meaning of stylizations such as changing a photograph to a line drawing[16], or fading it with time are taken directly from real life, neither are they wholly arbitrary[3]. Knowing the precise meaning of a face that appears as a drawing may require asking, but recognizing that those shown as full color realistic images are more “present” is clear.

Location. In the current implementation of *Collaboration-at-a-Glance*, the user can position the images of the other participants as he pleases in his window. This way, one can arrange a fairly large group of people so that one can easily see those who one is interested in. The information that *Collaboration-at-a-Glance* is designed to show best - presence and attention - can be conveyed while leaving the user free to rearrange the screen for the most convenience. In the current implementation, if, for example, Lindsey is corresponding with Arthur, and, on my screen, I move him from the right hand side to the left, Lindsey’s head will turn and follow the motion. (This may change in future implementations. See Section 3.1.)

Image features. The images for *Collaboration-at-a-Glance* are photographs, taken from a set of standard positions. The initial implementation used 7 frames per person; a later version used 27. The additional frames include some looking up or down, as well as additional sideways directions (see fig. x) All these frames all show the face in a neutral expression. Yet simply to show the person looking in “all” directions - or even in, say a fine-grained level arc around him/herself, would require an extensive database per person. Given that we are trying to establish, with a few easily distributed pictures, a set of frames

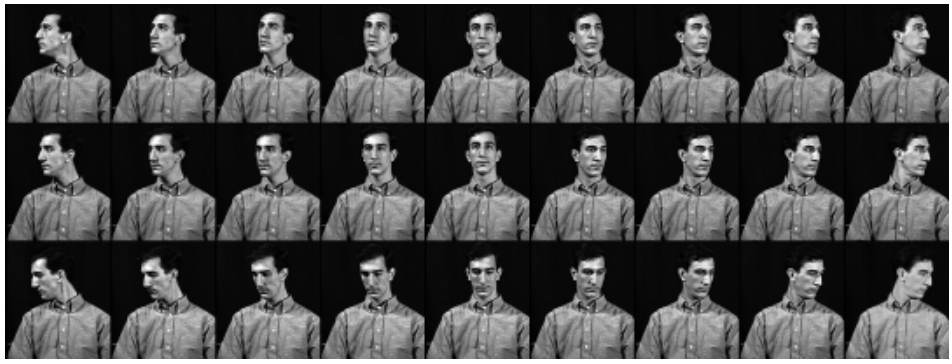

FIGURE 3. (right) An larger image set, allowing for finer resolution of gaze direction, plus vertical displacement.

that will let up portray someone looking in various directions, a very large database is not what we want.

3 Towards greater expression

Collaboration-at-a-Glance illustrates presence and attention: data that is already a part, albeit invisibly, of the electronic conversation. The next step in its development is to expand its communicative ability – what can it add to the expressiveness of the online discussion?

3.1 Location

With the current implementation, each user controls the layout shown in his or her own window, placing the individual images according to their relevance. The user cannot change the a particular image's style, for that is determined by the subject's presence; nor can the user determine which frame of an image is shown, for the direction of gaze is determined by the subject's attention[3]. Moving a picture from one place to another does not change the informational content of the window, for the heads turn to keep the subject of their gaze constant.

There is a trade-off between configurability and communication. Because the actions of the individual participants, rather than the preferences of the user, determine the direction each participant faces, information about attention is communicated to the user. Because the user controls the screen layout, no information can be communicated through location. What would be the likely advantages and disadvantages to giving up control of the layout and making location a means of communication?

In real life conversations, body language, including not only gesture, but also the spacing and positioning among the participants, plays an important role - one that is at times more telling than words. In a group conversation, those who are not speaking may still be communicating; they may nod their heads, or step slightly away from someone they disagree with (or who has been talking too long). A very real problem with online conversations is that there is no means to quietly indicate approval or doubt. With *Collaboration-at-a-Glance* allowing the participants to move about in a common space may provide a means to show solidarity, disagreement, or simple loss of interest.

Technically, the change would not be difficult. The server, which now keeps track of who is present and who is their focus of attention, would instead keep track of who is present and their location and orientation. A possible problem is disorientation; adding constraints, such as collision detection, should help alleviate this. The interesting question is how the ability to move about will be used.

In the real world, many conversational gestures and actions are performed as a way of controlling one's sensory input. One turns to face someone in order to see them; one moves closer to a conversation in order to hear it better. The fact that these actions are visible, and communicate information to others is the result, not the motivation. Virtual conversations do not have these sensory constraints: the participant can take in the whole scene at a glance; the text pours in, regardless of one's "location". The primary motivation

---

[3] I am using the term *user* to refer to the person looking at the conversation depicted in a window on his or her computer and *subject*s to refer to the people depicted in the windows. Each user is, of course, a subject on everyone else's screen.

for much conversational action seems to be missing. Does this mean that location cannot communicate meaning in electronic conversations?

Not necessarily. The key is the scale of the conversation - and the design of the interface. In a large electronic group where the conversation breaks off into many subtopics, filters, similar to those provided by the limits of the senses, can be useful. If, for example, a participant only receives the comments made by people who are within a certain radius, he is likely to move towards those whose discussion interests him. The result is a visible ebb and flow of discussions – an illustration of the dynamics within the group.

3.2 Expression

In *Collaboration-at-a-Glance* the expression on the face is neutral and unchanging.Yet in live conversations, variations in facial expression play an important role in the communication. People who see demonstrations of *Collaboration-at-a-Glance* often ask if a participant can make his or her image smile or look angry. How facial expression is interpreted in social situations and how people learn to produce the appropriate expression are questions that have been widely studied [4][11][12] and in this short paper I will not attempt to summarize this body of knowledge. Rather, I will discuss several of the problems that must be solved in order to design the interface so that a changing expression is intuitive and natural both to invoke and to interpret.

The three areas I will discuss are:

- **Invocation.** How does the user indicate a change of expression?
- **Feedback.** How does the user know what expression he or she is presenting?
- **Database.** What are the options for creating the necessary image database?

Invocation. The invocation method depends upon the range of the available expressions. If photographic stills are used, the set of expressions will be small and discrete. If synthesized images are used, the range and subtlety will be much greater.

For a limited expression domain a possible method for invoking expressions is verbal: the participant types the words “smile” or “angry” or “surprised”, and the expression on the representing image changes accordingly. Some evidence for the viability of this approach is found in the world of MUDs[7]. In this community, which is conversational but entirely text based, there has emerged a practice of indicating expressive actions verbally: if a user named Sal `types ":frowns"`, then all the other participants see `"Sal frowns"`.

In the MUD environment, the system does not need to know the meaning of the words - it simply broadcasts them. Describing one’s expression can itself be quite creative: Sal can `:smirk`, `:sneer`, or `:let the corners of his mouth drop slowly down til they reach his chest.` In a visual environment, the word would need to be translated to an image, so picking from a list - either verbal or pictorial - of available expressions would be more practical. The challenge then becomes to design the menu so that the user sees it as a quick and effective way to signal agreement, doubt, etc.

If a large and complex range of expressions are available, keeping the invocation method simple is important. Many existing methods, designed for animators, are far too complex. Their goal is precise control; here, the goal is to have an intuitive mapping from the input to the expression. A gesture based system would be good, one that, for example, mapped downward-sloping gestures to disapproval, back and forth motions to doubt, rapid upward motions to agreeable expressions.

Feedback. How does the user know how he or she appears to others? This is a problem even in real life: people are often unaware of what their expression reveals. Here, where even the minimal real-world feedback of muscle sensation – and other people's reactions – is missing, the users need to have some indication of the appearance they are presenting to the world. They need to know, for example, how long an expression lasts: if "smiling" is invoked, does the smiling expression stay until another expression is indicated? Or is it temporally bounded, fading to neutral after a few seconds? The feedback system can take the form of a "hidden camera" - an additional view that shows the user as he or she appears to others. Or it can be more abstract, such as a color shift or sound.

Underlying the design questions both of feedback and invocation are the basic challenges in the design of a first-person interface: creating a world in which one participates as oneself, where the technology both enables one to do things far beyond what is possible in the real world (such as sharing a space with people located hundreds or thousands of miles apart) and also limits one to a very constrained communication medium (such as requiring all action to be mappable to keyboard and mouse).

Image database. The two basic approaches to creating the image database are to record it photographically or to synthesize it.

The photographic approach is technically simpler and the images will generally resemble the person they are meant to represent. However, they are limited in range and number. For a system such as *Collaboration-at-a-Glance*, where every participant needs to have local access to the image database of the other participants, an immense set of frames per person (each expression, shown from all points of view) is not practical.

The synthetic approach [1][6][14][15][22][24], in which a photographic image is mapped onto a deformable 3D head model, while not yet feasible (the resulting images are likely to appear more grotesque than expressive) is quite promising. It would allow for a wide range of expressions, including subtle modulations. Yet, simply solving the technical problems of creating a truly human-like manipulable model and mapping the facial images onto it will not solve the problem of how to integrate it into a real-time conversational interface. Finding a way to control it so that the resulting expression is the one the user desired remains a difficult problem.

Regardless of how the database is generated, the important point is that the images should convey the expression that they are meant to show, not merely a physiologically correct, but perceptually and emotionally misleading version of it. The smile that is made by arranging the muscles in a 3D model – or the one produced by a participant while being photographed for an interface database – may be far from communicating the emotions of agreement or humor.

4 Summary

A conversation is a complex set of interactions. The text-only discussions that are currently found online emphasize the role of speaker; the listener is invisible. *Collaboration-at-a-Glance* creates a visual representation of a conversation, making it possible to quickly see who is participating and where is the center of attention. Yet there are many other non-verbal aspects to a conversation. This paper has discussed some of the issues involved in making *Collaboration-at-a-Glance* capable of communicating a greater range of expressive information These changes do not simply entail a

straightforward addition of features, but a deeper analysis of how the subtleties of conversational interaction can be transposed to an electronic environment.

References

[1] Agawa, Hiroshi; Xu, Gang; Nagashim, Yoshio; and Kishino, Fumio. 1990. Image analysis for face modeling and facial image reconstruction. Proceedings of SPIE: Visual Communications and Image Processing Vol 1360, 1184-1197.

[2] Brennan, S. 1982. Caricature generator. Master's Thesis, MIT.

[3] Brilliant, R. 1991.*Portraiture.* Cambridge, MA: Harvard University Press.

[4] Bull, Peter. 1990. What does gesture add to the spoken word. In *(*H. Barlow, C. Blakemore and M. Weston-Smith*,* eds.) *Images and Understanding*. Cambridge: Cambridge University Press, 108-121.

[5] Burke, Peter. 1993.*The Art of Conversation.* Ithaca: Cornell University Press.

[6] Choi, S.C.; Aizawa, K.;Harashima H. and Tsuyoshi, T. 1994Analysis and synthesis of facial image sequences in model-based image coding. In *IEEE Transactions on Circuits and Systmes for Video Technology.* Vol. 4. 257-275.

[7] Curtis, P. 1992. Mudding: social phenomena in text-based virtual realities. *Proceedings of the 1992 Conference on Directions and Implications of Advanced Computing.* Berkeley, May 1992.

[8] Donath, J. 1994. Casual collaboration. In *Proceedings of the International Conference on Multimedia Computing and Systems.* California: IEEE Computer Society Press.

[9] Dourish, P. and Bly, S. Portholes: Supporting Awareness in a Distributed Work Group. *Proceedings of ACM Conference on Human Factors in Computer Systems, CHI '92,* Monterey, CA.

[10] Goffman, E. 1959. *The Presentation of Self in Everyday Life.* New York: Doubleday.

[11] Gombrich, E.H. 1972. The mask and the face: the perception of physiognomic likeness in life and in art. In *Art, Perception and Reality.* Baltimore: The Johns Hopkins University Press, 1-46.

[12] Hochberg, J. 1978. Perception. 2nd ed. Englewood Cliffs: Prentice Hall.

[13] Hochberg, J. 1972. The representation of things and people. In *Art, Perception and Reality.* Baltimore: The Johns Hopkins University Press, 47-94.

[14] Koch, Reinhard. 1991. "Adaptation of a 3D Facial Mask to Human Faces in Videophone Sequences using Model Based Image Analysis". *Proceedings of the Picture Coding Symposium*, *Tokyo, Japan.* 285-288.

[15] Morishima, S. Aizawa, K. and Harashima, H. 1990. A real-time facial action image synthesis sytem driven by speech and text. In *SPIE Visual Communications and Image Processing*. Vol 1360, 1151-1158.

[16] Pearson, D.; Hanna, E.; and Martinez, K. 1990. Computer Generated Cartoons. In (H.Barlow, C. Blakemore, and M. Weston-Smith, eds.) *Images and Understanding*. Cambridge: Cambridge University Press, 46-60.

[17] Reid, E. 1991. *Electropolis: Communication and community on internet relay chat.* Thesis, Dept. of History, University of Melbourne.

[18] Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier.* MA: Addison-Wesley Pub. Co.

[19] Sproull, L. & Kiesler, S. 1991. *Connections: New Ways of Working in the Networked Organization.* Cambridge: MIT Press.

[20] Sproull, L. & Faraj, S. 1993. Atheism, sex, and databases: the net as a social technology. Forthcoming in (B. Kahin and J. Keller, eds.) *Public Access to the Internet.* Prentice-Hall.

[21] Sproull, L.; Subramani, R.; Walker, J.; and Kiesler, S. 1994. When the interface is a face. Unpublished manuscript.

[22] Takeuchi, A.and Nagao, K.. 1993. Communicative facial displays as a new conversational modality. In *Proceedings of Interchi '93.* ACM Press.

[23] Tufte, E. R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

[24] Waters, K. and Terzopoulos, D. 1992. The Computer Synthesis of Expressive Faces. Philosophical Transactions of the Royal Society B. 335, 87-93.