To appear in IFIP Working Conference on Engineering for Human-Computer Interaction (EHCI98), Heraklion, Crete, Greece September 14 - 18, 1998

Early experience with the mediaspace CoMedi

J. Coutaz, F. Bérard, E. Carraux, J. Crowley

CLIPS-IMAG
BP 53, 38041 Grenoble Cedex 9
Tel. +33 04 76 41 91 57, fax +33 04 76 44 66 75,
email: Joelle.coutaz@imag.fr

Abstract

Mediaspaces have been designed to facilitate informal communication and support group awareness while assuring privacy protection. However, low bandwidth communication is a source of undesirable discontinuities in such systems, resulting in a loss of peripheral awareness. In addition, privacy is often implemented as an accessibility matrix coupled to an all-or-nothing exposure of personal state. In this article, we describe CoMedi, a mediaspace prototype that addresses the problem of discontinuity and privacy in an original way: computer vision and speech recognition are used in conjunction to minimize visual discontinuities while supporting free movements in a room. Privacy is maintained by publication filters at the desired level of transparency.

Keywords

Computer mediated communication, mediaspace, privacy, group awareness, computer vision, face tracking, publication filter.

1 INTRODUCTION

The concept of «mediaspace» has been introduced in the early 90's as a means for facilitating informal communication and group awareness between spatially separated individuals. Informal communication denotes unplanned opportunistic encounters such as meeting someone by chance in the hall-way or glancing at someone through an opened door. Group awareness denotes a collective situated context for personal actions. It is grounded on the knowledge of the external world whether this knowledge is explicit, central and formal, or implicit, peripheral and informal, and whether it is useful now or in the future.

In typical mediaspace settings such as Cavecat (Mantei, 1991), Cruiser (Fish, 1992) and Montage (Tang, 1994), users can teleglance at a remote office, open a V-phone connection or maintain a permanent link with a distant shared location such as the commons. Although these services support informal communication and provide a global sense of a shared community (Dourish, 1992), they may also be used abusively and threaten privacy.

Access control to privacy may rely on social protocols as in the very first mediaspace developed at PARC (Stults, 1986) or it may use technical solutions as in Cruiser. Alternatively, access control may include a combination of both imperative and indicative controls as in Montage. Most solutions to disclosing privacy are two-fold: either the connection is permitted and an audio-video link is opened providing a full perceptual view on the distant location, or the connection is denied, and the distant visitor has no perceptual access to the remote site. The disclosure of private data is more complex than these simplistic binary solutions.

Another problem with mediaspaces is the restricted field of view on remote sites. As a result, peripheral awareness of distant people, objects, and events is lost. In addition, the static nature of the cameras induces extra articulatory tasks that interfere with the real world activity. For example, when V-phoning, users must keep their head within the field of the camera in order to be perceived by distant parties. Multiple views on remote sites improve the information bandwidth of a single static channel, but users have difficulties in linking the different views together (Gaver, 1993).

The mediaspace CoMedi has been developed as an answer to these concerns: group awareness and informal communication should be supported but privacy should be protected and sources of discontinuities should be avoided. In the next section, we describe CoMedi in details. Based on this early experience, we then present our research agenda for future development.

2 COMEDI

CoMedi (Communication and Mediaspace) is a mediaspace prototype that allows users to perform the following communication tasks: glance at someone, tele-visit a location using multiple forms of camera remote control, V-phone with someone while moving around in the office using a video tracking system, and publish private state variables through publication filters. Users are able to control the system using speech when they can't reach the mouse and the keyboard: glancing, opening a V-phone connection, etc., can be performed using either speech or direct manipulation.

2.1 The overall structure of the CoMedi user interface

The graphical user interface of CoMedi is structured into three functional parts: a menu bar, a porthole, and a control panel (see Figures 1 and 2).

The menu bar

At the top of the screen, the menu bar groups together the least frequent tasks. These include setting the access rights and the publication of sensitive data. When opening the *accessibility matrix*, the user can express the types of connection each distant user is allowed to issue. For example, the user can authorize good friends to both tele-visit his site and call him through the V-phone facility. On the other hand, less priviledged colleagues may not be allowed to tele-visit his place. Publication of sensitive data, which forms an original feature of CoMedi is discussed in 2.2.

The fisheye porthole

In the center of the screen, a fisheye porthole supports group awareness. A slot in the porthole corresponds either to a remote user or to a group of users. An individual slot displays the personal information that the remote user has accepted to make observable through publication filters. When opening a collective slot, the current porthole is replaced with the porthole of the corresponding group.

As shown in Figure 1, the porthole may have the shape of an amphitheatre where every slot is of equal size. When the fisheye feature is on, selecting a slot, using either the mouse or a spoken command, provokes an animated distorsion of the porthole that brings the selected slot into the center with progressive enlargement. For example, in Figure 2, Eric has selected Joëlle after having tuned the zoom factor using the slider below the porthole.

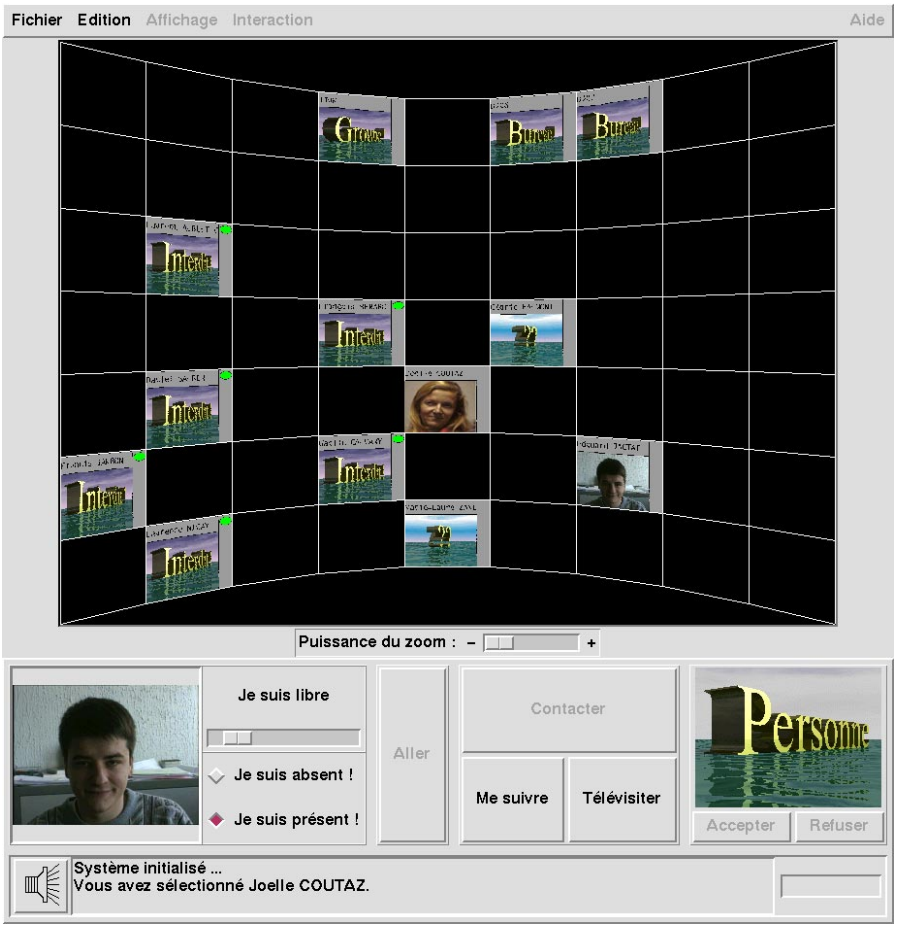

Figure 1 The graphical user interface of CoMedi is structured into three functional parts: at the top, a menu bar for non frequent tasks. In the center, a porthole that supports group awareness. At the bottom, a control panel for frequent tasks.

The motivation for a fisheye porthole is three-fold: it supports lightweight glancing, it provides detailed rendition in context, and it promotes scalability.

- At the opposite of most mediaspaces, glancing does not require any explicit action from the user except looking at the porthole.
- A selected slot denotes the current focus of attention. It is enlarged in a way that details about the remote activity are revealed at the appropriate level of

granularity. Meanwhile, the other slots, which denote peripheral attention, are shrinked but still observable to convey information about on going activities at remote sites.

- The porthole technique can accomodate a large number of users or group of users (see the 63 slots in Figure 1). For an even larger number of users, the slots may be too small to be discernable. If so, the porthole could be augmented with holophrastic techniques. For example, a number of slots would be used to synthesize the activity of multiple slots and indicate relevant state changes at remote sites. When selected, an aggregate slot would expand in place and reveal details progressively about the slot parts.

The control panel

At the bottom of the screen, a panel is dedicated to the most frequent control tasks: requesting or closing a remote connection, accepting or rejecting connection requests, checking personal sensitive data, and starting or stopping the tracking video system. Checking sensitive data is discussed in 2.2. The tracking system is presented in 2.3.

With regard to connection monitoring, CoMedi combines both imperative and indicative access controls: an authorized distant user can send a connection request using the Contact or the Tele-visit buttons (see the bottom right side of the control panel). This permission has been set up through the accessibilty matrix discussed above. Looking at the porthole, the caller can also check whether the distant user is currently available. Although he has the right to open a connection, the user may postpone the call based on social cues.

When contacted, the receiver can see the image of the caller (in the right most part of the control panel) as well as a pop-up timing band. He can accept or reject the request on the fly or ignore it. When the time out has elapsed, the connection returns to the idle state.

In the next sections, we present the original contributions of CoMedi for supporting privacy and minimizing discontinuities.

2.2 Supporting privacy through publication filters

In CoMedi, personal data include a business card and sensitive data rendered through publication filters.

Sensitive data

Sensitive data model the private space, for example, the fact that the user is currently in his office reading E-mail. Clearly, the observability of sensitive data is relevant to group awareness but may conflict with privacy. Our concept of *published observability* allows designers to reason about this antagonism (Salber,

1995). Sensitive data is made observable only if its owner has authorized its publication.

Figure 2 The porthole when the fisheye feature is activated. At the bottom left, according to the reflexivity property, the user can check and modify sensitive data. Here, Eric is using Venitian blinds to filter his private video scene.

In the current implementation of CoMedi, the types of sensitive data are the following:

- the absence or presence,
- the level of availability (available, busy, very busy, do not disturb),

- the audio scene (captured by the local microphone),
- the video scene (captured by the local camera).

The publication is set-up for each type of sensitive data through the top level menu bar. At the opposite of the accessibility matrix used for access control, the publication matrix is the same for every remote user. This design choice is motivated by implementation simplification only. In the example of Figure 2, Joëlle has authorized the publication of her video scene, but not her presence nor her level of availibility.

Exporting private sensitive data is one thing, remembering what is currently exported about oneself is a second thing.

Checking sensitive data: the reflexivity property

The capacity for the user to check the publication of his own sensitive data complies with the *reflexivity* property (Salber, 1995). The mirror image found in most tele-conferencing systems illustrates a simplistic case of reflexivity. CoMedi goes further: as shown at the bottom left of Figure 1, Eric can verify that he is currently publishing his private video scene as well as his presence and his level of availability. In addition, this portion of the control panel allows him to change the values of sensitive data (for example, switching the level of availibility from «busy»to «do not disturb»). The binary duality of publishing or hiding sensitive data does not however convey the subtlety of human social relations. We have introduced the concept of publication filter to satisfy this requirement (Salber, 1995).

Publication filter

A filter is a transformation function that applies to a set of published sensitive data. As shown in Figure 2, Eric is using a «Venitian blinds» filter that hides his private video space partially. Other filters such as posters are also available. In Figure 1, we observe that most users are hidden behind a poster. Setting a publication filter is performed through a pop-up menu that offers the list of available filters. The menu appears when clicking on the private video scene of the control panel.

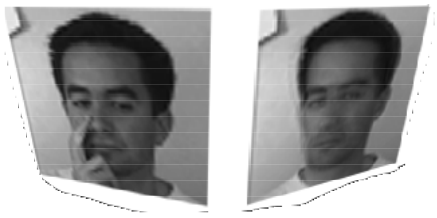

Figure 3 Eigen-space filtering for private video space. On the left, the source image; on the right the source image cleaned up through an eigen space filter.

Other techniques for filtering private video scenes have been developed recently: a low resolution image as in the Nynex Porthole (Lee, 1997), a ghost that denotes moving entities over time while blurring the entity itself (Hudson, 1996).

In (Coutaz, 1997), we present an innovative filter for video-based sensitive data using principal component analysis. Source images are coded as their coordinates in an N-dimensional orthogonal space: an eigen-space. This space is defined as the principal components computed from a set of representative publishable video images. The received image is rebuilt in real time on distant sites using a linear combination of the basis images. As a result, features in a source image that would not appear in the basis images are not reconstructed.

For example, in Figure 3, the source image showing François picking his nose is cleaned up through an eigen-space filter to produce a socially acceptable picture. Similarly any person appearing in the background would not be published to distant observers if not present in the basis images.

2.3 Minimizing discontinuities

Discontinuities in Computer Mediated Communication (CMC) is an opened problem. It covers multiple forms of disruptions primarily due to low perceptual bandwidth (Sellen, 1995). In CoMedi, we have investigated several ways of minimizing discontinuities for the two most relevant tasks in CMC: tele-visit and audio-video communication with distant users. We have addressed the visual dimension of discontinuity using computer vision and image processing in conjunction with speech.

The computer vision tracker

The computer vision tracker developed for CoMedi uses a pan-tilt-zoom color camera. As shown in Figure 4, it is based on an architecture in which a supervisor activates and coordinates three visual complementary processes: eye blink, color histogram, and cross-correlation (Coutaz, 1996; Crowley, 1997).

Eye blink detection is based on the difference of successive images. If the eyes happened to be closed in one of the two images, two small roundish regions appear over the eyes where the difference is significant. Eye blink detection is computationally inexpensive and can handle a wide range of lightning conditions.

The head position estimation provided by eye blink detection is used to calibrate the color histogram from a region of the image (close to the eyes) that contains skin colored pixels. Color histogram is computationally cheap but sensitive to camera noise resulting in jittering. To achieve accuracy, correlation tracking is used.

Cross-correlation operates by comparing a reference template (e.g., piece of an eyebrow) to an image neighborhood at each position within a search region. Cross

correlation is very accurate but it is unable to track head rotation. This problem is solved using the cooperation of the two other complementary detection techniques.

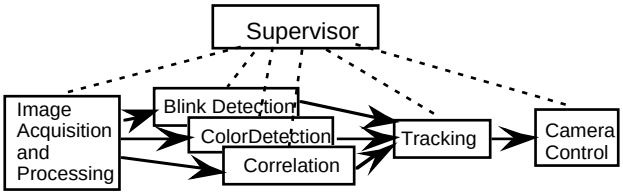

Figure 4 A Supervisory controller selects and controls the sequencing of visual processes. Dotted lines denote activation by the supervisor. Arrows express the main stream data flow.

In order to support cooperation, the output provided by each visual process is normalized and formalized: each process returns its estimation of the head position, a precision and a confidence factor. Figure 5 illustrates the cooperation of the three visual processes.

When tracking confidence is low, the supervisor runs blink detection to look for a face (eye blink is fast and does not need initialisation). When blink is detected, a color histogram is initialised, a correlation template is stored for each eye. As long as the tracking CF remains high, correlation is used to track the eyes (correlation is fast and precise). When a tracking CF with a low value is obtained, correlation tracking has failed, the color histogram is used to recover the face (histogram always returns a result). If the tracking CF is high again, the correlation template is re-initialised at an eye position estimated from the face position and the correlation is run again. If, on the other hand, the tracking CF drops below a threshold, the supervisor draws upon the eye blink detector.

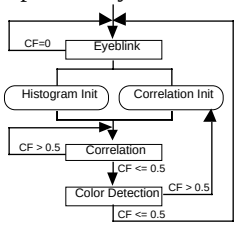

Figure 5 Cooperation of the perceptual processes techniques.

Minimizing discontinuities with the computer vision tracker

In CoMedi, the computer vision tracker is used to control the movement of the local camera. It can also be used to control the field of view of a remote camera as in Gaver's virtual window (Gaver, 1995).

To start/stop the local tracker, the user can utter the sentence «follow me»/«stop following me» or click the «follow me» button in the control panel (in French: «me suivre»). Local tracking avoids users to check their location within the field of view of the camera. The migration to the system of counterproductive articulatory tasks minimizes discontinuities in the communication process.

In addition, local tracking opens the way to a new kind of interactive user: a user who is not wired to a terminal, a user free to move. Because the tracking system is based on multiple visual processes, it has the potential to smoothly shift from the head target (i.e., talking head mode) to the hand pointing at a new object of interest (e.g., the drawing on the blackboard the users are currently talking about). As users talk, it is possible for them to move around while the local camera adjusts the field of view dynamically.

To start/stop the virtual window, the user can utter the sentence «tele-visit»/«stop tele-visiting» or click the «tele-visit» button in the control panel (in French: «télévisiter»). As users talk, they can telecontrol the remote camera by moving their head and ajust the remote field of view according to their needs. Because users can explore distant sites under their own control just like they would do in front of a physical window, Gaver has demonstrated a decrease in visual discontinuities. On the other hand, the virtual window technique offers one single view at a time. As observed by Gaver et al., «one [view] is not enough» (Gaver, 1993). We have developed Fovea, a technique based on image processing, to address this problem.

Minimizing discontinuities with Fovea

The motivation for Fovea is to avoid the visual discontinuity due to the split screen solution adopted in Extra-Eyes (Yamaashi, 1996). In Extra-eyes, a low resolution camera provides the observer with a wide angle view of the remote site. A detailed view of the focus of interest is obtained at a high resolution in a second window. Although a rectangle is drawn in the wide angle view to show the location of the detailed view, the user has to consolidate the visual discontinuity between the two views.

Fovea fuses multiple views into a single image. It is inspired by foveal and peripheral architecture of the human visual system. As shown in Figure 6, the image received at a distance results from the combination of a low resolution image (the periphery) with a high resolution image (the fovea). In its current implementation, the location of the fovea in the remote scene is controlled with the mouse (but the head tracker described above could be used as well).

As the user explores the remote site, he may find something of interest, for example a postcard pinned on the wall. As shown on the right side of Figure 6, the zoom facility of the fovea can be used to obtain information at the right level of detail without loosing the context. From preliminary user studies where six users

were asked to find random targets, users were more efficient with a rectangular fovea than with Extra-Eyes but they were more efficient with Extra-Eyes than with the circle shape fovea. On the other hand, all of them prefered the circular fovea. These preliminary results need additional experimentation to confirm our findings.

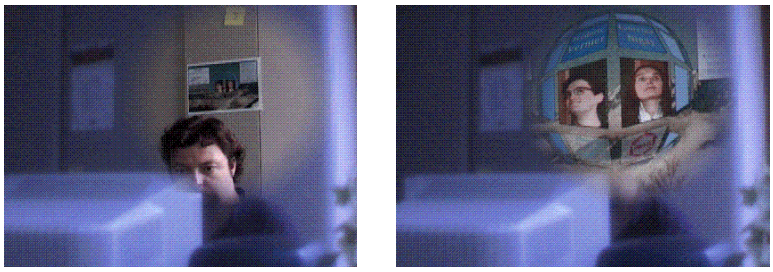

Figure 6 On the left, a video image of a remote site that combines a low resolution picture with a high resolution fovea . It is currently pointing at a picture on the wall. On the right, the video image when using a zoomed fovea.

Minimizing discontinuities using speech

The motivation for speech is to support mobility: as the user moves away from the mouse and the keyboard, he can still control the mediaspace using speech.

Spoken control in the context of audio communications may generate confusions. In particular, the speech recognition system needs to identify whether the user is talking to a distant partner or uttering a system command. In general, speech systems can be used in a «push to talk» mode or in a continuous recogition mode. The push to talk approach increases system robustness but induces an extra articulatory task. The continuous mode is free from extra task but computationaly more expensive and less robust.

In CoMedi, we use the speech system in the continuous recognition mode to eliminate the push button additional task but we prevent the system from listening by hiding the microphone: The HF receiver/transmitter is placed in the user's pocket while the microphone itself is clipped on the wristwatch (not on the shirt collar!). In general, users keep their left hand far from the mouth. To talk to the system, the user makes the same gesture as for reading the watch. Although the naturalness of the setting has not been tested formally, early experience indicates easy acceptance.

3 LESSONS LEARNED AND RESEARCH AGENDA

The lessons learned from our early experience with CoMedi are both technical and human centered.

3.1 Technical aspects: CoMedi is a concept demonstrator

CoMedi is implemented according to the PAC* architecture model (Calvary, 1997). Its functional core, which maintains the data base of users, is an active data structure implemented as a GroupKit environment (Roseman, 1992). The Interaction and the Presentation components host the modality interpreters: the speech recognition system (ViaVoice from IBM), the computer vision tracker, and the graphical abstract machine Tk. These interpreters are all encapsulated in Tcl providing a uniform view to the Dialogue Component. The Dialogue Component is refined in terms of PAC agents according to the PAC style (see Coutaz, 1997b for a more detailed description of this CoMedi component).

For efficiency, modality interpreters are distributed over multiple processors: a PC runs the speech recognition system while a SGI Indy is dedicated to computer vision and a second one runs the rest of the local computations. As a result, every CoMedi user needs 3 workstations to be part of the mediaspace. Clearly, CoMedi is a concept demonstrator. Its computational cost is too prohibitive for an effective large scale use.

In addition, the absence of guaranteed bandwidth makes impossible tightly coupled interaction over the Ethernet. For example, the remote control of cameras as in Fovea or the virtual window, is difficult to achieve with unstable response times.

Based on this early experience, we have developed a lightweight version of CoMedi for a realistic use in the IMAG community. So far, CoMedi Light does not include any computer vision tracker nor speech recognition. Interpersonal communication is limited to glances and Post-It messages. On the other hand, CoMedi Light is portable on standard platforms (MacOs, Irix and Windows NT) connected over the Ethernet; its efficient use of resources makes it possible non-stop background execution. Figure 7 shows the actual graphical user interface of CoMedi Light implemented in Java. It has been in use for one month only between 15 volunteers who daily meet face-to-face (publication filters have not been installed yet). Although it is too early to report on CoMedi Light as a social object, we have already observed interesting social phenomena

Figure 7 The graphical user interface of CoMedi Light.

3.2 Human centered aspects

The technical limitations observed for both CoMedi and CoMedi Light add new items to our research agenda: balance of the equipment, camera full coverage, filter refinement, audio and video scenes analysis.

Balance of the equipment

The disparity between mediaspace workstations may have undesirable effects on social behavior. For example, 5 CoMedi Light users, who did not have a camera, would not launch the mediaspace. They said they would see other people without being observed. In this case, the lack of reciprocity due to unbalanced equipment induced an undesirable sense of voyeurism. Unbalanced equipment, whose increase follows the advent of mobile and ubiquitous computing, is an important issue to address in order to guarantee social fairness in CMC.

Camera full coverage

Although Fovea and the virtual window support tele-exploration, it is still difficult for remote observers to identify which part of the distant place is currently seen by the camera. There are two reasons for this: a single camera is not enough to scan 360º; second, people take possession of their camera: they keep changing its location. As a result, users cannot develop a stable model of remote sites.

Camera full coverage can be addressed through the cooperation of multiple cameras. In addition, each camera should be able to dynamically compute its own location within the scene. With this information, it is then possible to provide users with an additional picture that makes concrete the current field of view of the camera within the remote scene.

A filter algebra

So far, filters work independently and use one single sensitive information at a time. For example, video filters apply to the whole scene as opposed to a patchwork composition. As an example of composition, a poster may hide the latest results written on the blackboard, the user's face may be cleaned up through an eigen-space filter, and the visitors may be blurred with a low resolution picture. As the user and the visitors move around, filters may overlap: what is the resulting image?

Fusing multiple sensitive data into a meaningful synthesized representation is also another way to go. For example, combining keystrokes, mouse actions, differences of audio and video images into a single a level of activity. But the resulting combination of filters should not be reversible in order to prevent any form of maleficence.

We plan to develop a filter algebra to reason about publication filtering.

4 CONCLUSION

In this article, we have described CoMedi, a prototype mediaspace, that addresses the problem of discontinuity and privacy in an original way. This heavy weight concept demonstrator has opened the way to the development of a portable version for effective use by a large community of users: Comedi Light. The services of CoMedi Light will be incrementally augmented along the lines of our research agenda.

computer vision and speech recognition are used in conjunction to minimize visual discontinuities while supporting free movements in a room. Privacy is maintained by publication filters at the desired level of transparency.

5 ACKNOWLEDGMENT

This work is being supported by France Telecom-CNET.

6 REFERENCES

Calvary, G., Coutaz, J. and L. Nigay (1997) From Single-User Architectural Design to PAC*: a Generic Software Architecture Model for CSCW. In *Proceedings of CHI 97*, ACM publ., pp. 242-249.

Coutaz, J. , Crowley, J. L. and Bérard, F. (1996) Coordination of Perceptual Processes for Computer Mediated Communication, *in proceedings of the second International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society Press, pp. 106-111.

Coutaz, J., Crowley, J. and F. Bérard (1997) Eigen space Coding as a Means to Support Privacy in Computer Mediated Communication. *In Proceedings of INTERACT'97*, Chapman & Hall Publ.

Coutaz, J. (1997b) PAC-ing the Software Architecture of your User Interface. *DSV-IS'97, 4th Eurographics Workshop on Design, Specification and Verification of Interactive Systems*, 1997, Springer Verlag Publ., pp. 15-32.

Crowley, J. L. and F. Bérard (1997) Multi-Modal Tracking of Faces for Video Communications, *IEEE Conference on Computer Vision and Pattern Recognition CVPR '97*, Puerto Rico.

Dourish, P. and Bly, S.A. (1992) Portholes : Supporting Awareness in a Distributed Work Group, in *proceedings of the CHI'92 Conference on Human Factors in Computing Systems*, pp. 541-547.

Fish, R.S., Kraut, R.E., Root, R.W., and Rice R.E. (1992) Evaluating Video as a Technology for Informal Communications, in *proceedings of the CHI'92 Conference on Human Factors in Computing Systems*, pp. 37-47.

Gaver, W., Sellen, A., Heath, C. and P. Luff (1993) One is not Enough: Multiple Views on a Media Space, *Proc. INTERCHI'93*, ACM Publ., pp. 335-341.

Gaver, W., Smets, G. and K. Overbeeke (1995) A Virtual Window on a Media Space, *Proc. CHI'95*, ACM publ., pp. 257-264.

Hudson, S. & Smith, I. (1996) Techniques for Addressing Fundamental Privacy and Disruption Tradeoffs in Awareness Support Systems. In Proc. CSCW'96, ACM Press, pp. 248-257.

Lee, A., Girgensohn, A. and Schlueter K. (1997) NYNEX Portholes: Initial User Reactions and Redesign Implications. *In Proc. GROUP'97, International Conf. on Supporting Group work*, ACM Publ.

Mantei, M., Backer, R.M., Sellen, A., Buxton, W., Milligan, T. and Wellman B. (1991) Experiences in the use of a Media Space, in *proceedings of the CHI'91 Conference on Human Factors in Computing Systems*, pp. 203-208.

Roseman, M. and Greenberg, S. (1992) GROUPKIT: A groupware Toolkit for Building Real-Time Conferencing Applications, *in Proc. CSCW'92*, ACM Conference on CSCW, pp. 43-50.

Salber, D. (1995) *De l'interaction Homme-Machine individuelle aux systèmes multi-utilisateurs: l'exemple de la communication homme-homme médiatisée*, Thèse de doctorat de l'Université Joseph Fourier, Grenoble.

Sellen, A. (1995) Remote Conversations: The Effects of Mediating Talk with Technology, in *Human Computer Interaction*, Lawrence Erlbaum Publ., Vol. 10(4), pp. 401-444.

Stults, R. (1986) *MediaSpace*, rapport technique Xerox PARC.

Tang, J.C. and Rua, M. (1994) Montage: Providing Teleproximity for Distributed Groups, in *Proc. of the Conference on Computer Human Interaction (CHI'94)*, pp. 37-43.

Yamaashi, K., Cooperstock, J., Narine, T. and Buxton, W. (1996) Beating the limitations of Camera-Monitor Mediated Telepresence with Extra Eyes, In *proc. CHI'96*.

7 BIOGRAPHY