Coordination of perceptual processes for Computer Mediated Communication

Joëlle Coutaz, François Bérard
CLIPS-IMAG, BP 53
38041 Grenoble cedex 9 France
{Joelle.Coutaz, Francois.Berard}@imag.fr

James L. Crowley
GRAVIR-IMAG
I.N.P. Grenoble, 46 Ave Félix Viallet
38031 Grenoble France
jlc@imag.fr

Abstract

In Computer Mediated Communication such as desk-top video conferencing, static video cameras provide a restricted field of view of remote sites. The effective field of view can be enlarged, while maintaining the user's freedom of movement, by slaving a remote controlled camera to movements of the user's head. This paper concerns techniques for tracking of faces. We demonstrate that robustness and reliability can be increased by combining multiple perceptual processes such as eye blink detection, skin color histogram and cross correlation, that adapt to a variety of operating conditions. We illustrate our technique with CoMedi, a media-space currently under development.

Key-words: face tracking, data fusion, integration of visual processes, media space.

1. Introduction

Computer Mediated Communication (or CMC) covers multiple forms of person-to-person communication supported by a computer network infrastructure [8]. CMC may occur asynchronously as with E-mail or synchronously as in desk-top video conferencing. In typical synchronous CMC settings, video cameras provide a restricted field of view of remote sites. As a result, peripheral awareness of distant people, objects, and events is lost. In addition, the static nature of current CMC apparatus induces extra articulatory tasks that interfere with the real world activity. For example, 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 [6].

Multiple views can be provided without spatial discontinuities by using a "Virtual Window' in which a remote camera is slaved to a person's head movements [7]. However, the lack of robustness of the vision technique developed in [7] to track the user's head (difference image, head assumed to be upright in the image) imposes restrictions on the user. Typically, once a satisfactory point of view has been reached, users dare not move further. It has been demonstrated that in CMC, robustness and performance have a strong impact on user's behavior and system acceptance [10].

The central message of this paper is that robustness and performance can be increased by combining multiple perceptual processes that adapt to a variety of operating conditions. We illustrate this message with CoMedi, a media-space under development for experiments with both social and technical aspects of CMC. We first introduce the CoMedi system. We then present the architecture used as the framework for integrating multiple perceptual processes followed by the description of each of the perceptual processes implemented in CoMedi. We close the discussion with an example of cooperation of visual processes based on their mutual merits and limitations for a robust and efficient face tracking.

2. CoMedi: a media-space prototype

The infrastructure of a media-space is similar to that of a desktop video-conferencing system. Whereas tele-conferencing is formal and pre-planned, media-spaces are intended to support opportunistic encounters such as meeting someone by chance in the hall-way or glancing at someone through an opened door. CoMedi (Communication and Media-space) is a media-space prototype that allows users to perform the following tasks: glance at someone (a short duration video connection with a distant party), tele-visit (e.g., Vphone and exploration of a distant location such as a public area using a virtual window), and protection of the private space (e.g., closing the door for everybody except for one's best friend). This system is currently being tested on four Silicon Graphics Indy workstations communicating over the Ethernet. Each workstation is equipped with a Canon VCC1 motorized pan-tilt-zoom color camera, and microphones placed on the four corners of the computer monitor.

CoMedi is enriched with multiple active and cooperative perceptual systems. Three visual processes (eye blink, color histogram, cross-correlation, and one audio process) are currently used in cooperation to detect and track media-space occupants. Because the 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). Similarly, as users talk, it is possible for them to move around while the local camera

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996

adjusts the field of view dynamically.

3. A synchronous ensemble of reactive visual Processes

The tracking system described in this paper is based on an architecture in which a supervisor activates and coordinates perceptual processes. We call such an architecture a Synchronous Ensemble of Reactive Visual Processes (SERVP). This architecture has first been developed in the context of robotics [3] and surveillance tracking [5].

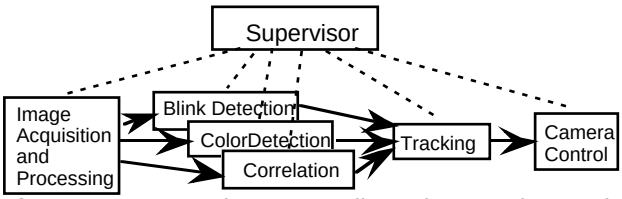

Figure 1. A Supervisory controller selects and controls the sequencing of perceptual processes. Multiple processes can be active at the same time. Dotted lines denote activation by the supervisor. Arrows express the main stream data flow.

3.1. The SERVP architecture

As shown in Figure 1, the core component of the SERVP architecture is a supervisor which drives the system in a cycle composed of 5 phases.

1) Process selection: Based on the currently specified task and system state, the supervisor enables and disables visual processes. The order of processes is determined and priority is assigned for resolving conflicts in device commands and time allocation. In CoMedi, the processes selection is organised as a network of states.

2) Virtual sensors activation. Based on the currently active set of processes, it is possible to anticipate certain image processing requirements such as resolution reduction. The supervisor selects these common "low-level" operations (or virtual sensors) which produce results shared by the visual processes. In CoMedi, a virtual sensor produce low resolution black and white copies of the current high resolution color image.

3) Visual processes activation. Selected visual processes are activated in sequence by the supervisor. Processes are defined as a transformation from the current image to an observation vector. The observation is communicated to the supervisor and to a fusion and tracking process. In the case of CoMedi, the observation is composed of the position and size of the face. Observations are time-stamped and completed with a covariance matrix to express precision and a confidence factor to express the success of the detection.

4) Fusion and tracking. Based on the results provided by the active visual processes, the fusion and tracking process maintains an estimate of the center point and the size of the face using a form of recursive estimator described below.

5) Camera control. The current estimate of position and size of the face provides a reference signal to a PD controller for pan tilt and zoom of an RS232 controlled camera.

3.2. Fusion and integration in a recursive estimator

The use of estimation theory for tracking and for fusion of information in computer vision and robotics is well established [1], [2]. For our face tracker, we use a zero-th order Kalman filter to maintain estimates of the center position of the face (i, j) and the vertical and horizontal size of the face, (h, v). We estimate the horizontal and vertical size of the face as two parameters because the aspect ratio of the face can change with rotations. Thus the state vector, X, maintained by the tracking process has four components, (i, j, h, v), measured in pixels.

Fusion of perceptual information computed by the visual processes is made possible by an explicit estimate of the precision and confidence of each observation [5]. Precision is represented by a covariance matrix, $\mathbf{C}_X$. The state vector X and its covariance $\mathbf{C}_X$ are estimated in a recursive process composed of three phases: predict, match and update, illustrated in figure 2. In the text that follows, we will use the convention that "^" denotes estimated quantities, and "*" denotes predicted quantities.

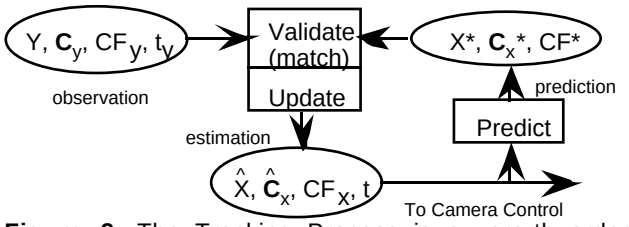

Figure 2. The Tracking Process is a zero-th order recursive estimator for position and size.

Predict: Each visual process provides an observation Y of the state vector (or a subset) accompanied by a time stamp, t_y, a covariance matrix, $\mathbf{C}_y$, and a confidence factor, CF_y. In order to update the estimate with the observation we must first predict the value of the estimate at the time of the observation. This is the task of the prediction phase .

The time stamp, t_y, is used to compute a time step, Δt, from the time the estimation was last updated, t_X:

$$\Delta t = t_y - t_X$$

Movements of the subject between observations are unpredictable. Thus we make no attempt to estimate temporal derivatives. The prediction of the state vector, X^* at time t_y, is simply the last updated estimate of X at time t_X, $\hat{X}$:

$$X^* := \hat{X}$$

The covariance, on the other hand, does depend on the time step. The uncertainty in position of the subject is a

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996

quadratic estimate which grows as the square of the time step. This growth is captured in a 4x4 matrix **W**, whose terms give the loss in growth in uncertainty of each component of $\mathbf{C}_X$ as a function of seconds-squared. Thus the covariance is updated as:

$$\mathbf{C}^*_X := \hat{\mathbf{C}}_X + \Delta t^2 \, \mathbf{W}$$

The coefficients of **W** are calibrated from a sequence of position and size estimates of a normal user. The confidence of the estimate is degraded by multiplication by a factor, α, which is less than 1, raised to the power of the time step.

$$CF_X^* = \hat{CF}_X \, \alpha^{\Delta t}$$

Match: The state vector, X^*, which is predicted for time t_y, is compared to the observed state vector, $\vec{Y}$, using a Mahalanobis distance, d. The Mahalanobis distance is the difference between the observation and the estimation, normalised by the sum of their covariances. This provides a validation gate which shrinks and grows with the precision with which the position and size of the face are known.

$$d^2 = \frac{1}{2}(X^*-\vec{Y})^T (\mathbf{C}^*_X+\mathbf{C}_y)^{-1} (X^*-\vec{Y})$$

If the Mahalanobis distance between the observation and the estimation exceeds a threshold, then the observation is rejected.

Update: An estimated vector and its covariance represent the first and second moments of a probability distribution. In statistical estimation theory, physical laws for combining moments are used to fuse an observation with an estimate. The covariance matrices provide the weights for a weighted average of the observation and estimation. Updating the estimate requires first computing a new estimated covariance matrix given by:

$$\hat{\mathbf{C}}_X := (\mathbf{C}^{*\,-1}_X + \mathbf{C}_y^{-1})^{-1}$$

The new estimated vector can then be computed as a weighted average:

$$\hat{X} := \hat{\mathbf{C}}_X (\mathbf{C}^{*\,-1}_X X^* + \mathbf{C}_y^{-1} \vec{Y})$$

This is mathematically equivalent to the more commonly used Kalman gain matrix.

3.3. Estimating the confidence of visual processes

Confidence (CF) is estimated as the probability that a successful detection was achieved. The probability of observations are computed using a sample set of correct detections, represented by a mean, $\vec{\mu}_S$, and covariance, $\mathbf{C}_S$, obtained during system set up. The confidence of each observation is computed by comparing observed vector, Y, to this pre-calibrated mean and covariance using a Gaussian density function. Confidence of an observation is combined with the confidence of the estimate as independent probabilities.

$$\hat{CF}_X = CF_X^* + CF_y - CF_X^* \, CF_y$$

4. Perceptual processes for detection

Robust continuously operating tracking can be obtained by driving the tracking process with several complementary detection processes. The tracking process then provides a reference with which individual processes can be re-initialised when their result becomes unreliable. This section describes processes for detecting faces using blinking, normalised color, cross-correlation and sound.

4.1. Detecting faces from blinking

A human must periodically blink to keep his eyes moist. Blinking is involuntary and fast. The fact that both eyes blink together provides a redundancy which permits blinking to be discriminated from other motions in the scene. The fact that the eyes are symmetrically positioned with a fixed separation provides a means to normalize the size and orientation of the head from the detection.

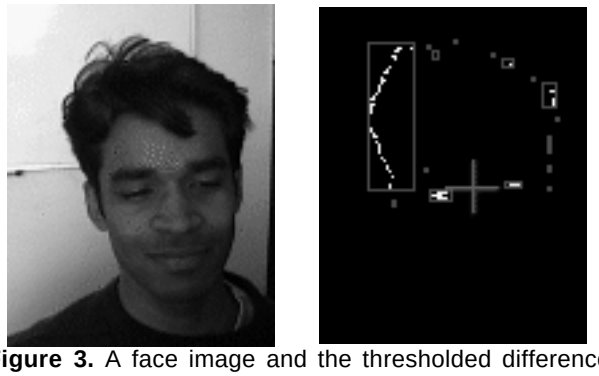

Figure 3. A face image and the thresholded difference with bounding boxes and face position.

Blink detection is based on the difference of successive images. The difference image generally contains a small boundary region around the outside of the head. If the eyes happened to be closed in one of the two images, there are also two small roundish regions over the eyes where the difference is significant, as shown in figure 3.

The difference image is thresholded, and a connected components algorithm is run on the thresholded image. A bounding box is computed for each connected component. Candidate regions for an eye are selected based on horizontal and vertical size of the bounding box. Candidate regions are then paired and tested for a small vertical displacement and an appropriate horizontal separation. When this configuration of two small bounding boxes is detected, a pair of blinking eyes is hypothesized. The position in the image is determined from the center of the line between the bounding boxes. The distance to the face is measured from the separation. This provides the size of a window which is used to extract the face from the image. This simple technique has proven quite reliable for determining the position and size of faces [9].

Blink detection produces a vector of 8 components:

v_l, h_l Vertical and horizontal sizes of left rectangle

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996

v_r, h_r Vertical and horizontal sizes of right rectangle

v_s, h_s vertical and horizontal separations of the mid-points of the rectangles

i , j horizontal and vertical components of mid-point between rectangles.

The midpoint between the rectangles is used as the observation of the position of the face, [i, j]. The horizontal and vertical size of the face can be computed from a scale factor S, taken as the ratio of the distance between the eyes, to a precalibrated "ideal" value. The scale factor is then multiplied by the size which corresponds to the ideal separation. The covariance matrix for position $\mathbf{C}_b$ is given as a constant which is calibrated during system set-up.

The confidence of a blink detection, CF_b, is the resemblance of the eight parameters to an ideal prototype, P_{blink}, and its covariance $\mathbf{C}_b$. This prototype is computed a priori by recording a large number of blink detections and removing any false detection by hand.

4.2. Detecting the colour of skin

Color histograms have been used in image processing for decades, particularly for segmenting multi-spectral satellite images, and medical images. In the early 1990's Swain and Ballard [11] showed that the intersection of color histograms was a reliable means of recognizing colored objects. Unfortunately, their technique is sensitive to the color and intensity of the ambient light source. As demonstrated in [9], skin can be reliably detected by normalising the color vector by dividing out the luminance component.

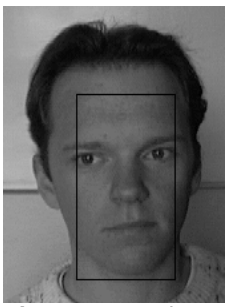

Figure 4a The color histogram is initialised from pixels in rectangle.

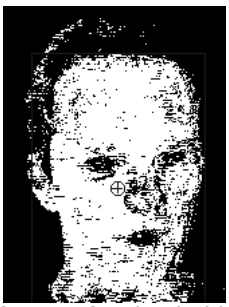

Figure 4b Thesholded probability of skin.

A 2-D joint histogram of the luminance normalised color components (r, g) can be computed from a patch of an image known to be a sample of skin. For color components (R, G, B):

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B}$$

The histogram of normalised color gives the number of occurrences for each normalised color pair (r, g). A normalised color histogram h(r, g) based on a sample of N pixels, gives the conditional probability of observing a color vector $\vec{C}$ = (r, g), given that the pixel is an image of skin, p($\vec{C}$ | skin). Using Bayes rule, we convert this to the conditional probability of skin given the color vector, p(skin | $\vec{C}$). This allows us to construct a probability image in which each pixel is replaced by the probability that it is the projection of skin. An example is shown in figure 4. A normalised color histogram is computed from the rectangle in figure 4a. The probability of skin is then computed from later images to produce the binary region shown in figure 4b. The center of gravity from the probability of skin gives the estimate of the position of the face. The bounding rectangle gives an estimate of size. A confidence factor is computed by comparing the detected bounding box to an ideal width and height, using a Normal probability law. The average width and height and the covariance matrix are obtained from a number test observations selected by hand.

4.3. Tracking with cross-correlation

Cross-correlation operates by comparing a reference template to an image neighborhood at each position within a search region. The reference template is a small neighborhood, W(m, n), of size M by N, of the image P(i, j) obtained during initialisation. The search region is estimated from the expected speed of the user's movements measured in pixels per frame. In subsequent images, the reference template is compared to each image neighborhood within the seach region. Our system includes comparisons by energy Normalised Cross-Correlation (NCC) and by Sum of Squared Differences (SSD) [4].

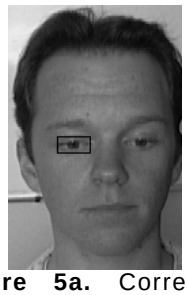

Figure 5a. Correlation template is taken from eye using blink detection.

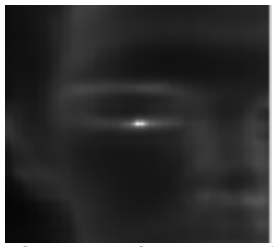

Figure 5b. Map of values from Sum of Squared Difference.

The estimated position of the target is determined by finding the position (i, j) at which the SSD measure is a closest to zero. The actual center position can be determined by adding the half size of the template to the corner position (i, j).

$$X_{cor} = (x, y) = (i, j) + (\frac{M}{2}, \frac{N}{2})$$

Figure 5 shows a typical map of the SSD values obtained when a template for the eye is convolved with a face. The local image of SSD values is inverted to provide the CF and covariance. The covariance of the detection is estimated from the second moment of the inverted SSD values. A sharp correlation peak gives a small covariance, while a larger correlation gives a larger

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996

spread in covariance. The confidence is estimated from the peak value of the inverted SSD. When this confidence measure drops below a threshold the cross-correlation process is halted or re-initialised.

4.4. Locating faces using speech utterances

Many modern computer workstations, such as the SGI Indy and the Macinstosh Quadra AV, are equipped with integrated digitizers for images and sound. It is possible to use the phase difference between pairs of microphones mounted on the workstation monitor to determine the angle to the speaking human. This permits an estimation of the location of the speaker with a precision on the order of a few centimeters in horizontal and vertical position. Distance from the monitor is less precise.

Relative time delays between the signals received by pairs of microphones are estimated using biased cross-correlation. By placing the microphones in a "T" configuration (top center and corners, bottom center) in order to improve the horizontal precision. A signal received the center microphone is taken as a reference and time delays between this signal and the signals from the other microphones are estimated. Each time delay corresponds to a hyperbolic surface, which may be approximated as a plane. The nominal face position produces a nominal time shift vector, $\vec{\Phi}$. The difference, $\Delta\vec{\Phi}$, between the nominal time shift and the measured time shift is multiplied by Jacobian matrix to give the change in position and size of the image in pixels.

$$\vec{Y} = \mathbf{J}\ \Delta\vec{\Phi}$$

The covariance of these displacements is estimated by calibration experiments. The confidence is given by the normalised cross correlation score. Reliable position estimation from speech has only recently been demonstrated in our project, and has not yet been integrated into the complete system.

5. Coordination of perceptual processes

As shown in Figure 6, the perceptual processes of eye blink detection, color histogram matching, and correlation tracking are complementary. Each process fails under different circumstances, and produces a different precision for a different computational cost. For example, eye blink is relatively inexpensive in cost and gives a precise localisation when it works, which is approximately once every 30 seconds. Thus eye blink is ideal for initialising, and re-initialising, the other tracking processes. Correlation tracking of the eyes is extremely fast when limited to a small search region and produces a precise result. However, experience shows that correlation will sometimes loose its track when the user turns his head more than about 15 degrees or makes a movement which is too sudden. In some cases, correlation can be recovered by enlarging the search region, but if this fails, another tracking mode is required.

Color histogram matching almost always produces a result but tends to have an uncertainty of a few pixels and must be periodically re-initialised to compensate for changes in ambient light, or differences in skin color of different users. In our early experiments with this technique, a cooperative user presented his face or hand to the camera to initialize the histogram in less than a second. In our latest system, the color sample is captured automatically whenever eye blink has been detected with a sufficient confidence.

If computing cost were not a constraint, all visual processes would be run at each cycle of the system. In order to achieve (soft) real time execution, it is necessary to select a subset of processes to run in each cycle. This is made possible by the existence of a confidence factor .

In CoMedi, the control logic for the supervisor is defined by a finite state machine. A current doctoral thesis in our group is investigating techniques to automatically generate such control graphs. In the mean time, we design control graphs by hand. At the time of writing of this paper, we obtain quite reliable tracking (the tracking precision, as measured by the covariance remains under 2 pixels) with the following states:

State 1) 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 and a correlation template is stored for each eye. The supervisor then shifts to state 2.

State 2) 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 (< 0.5) is obtained, correlation tracking has failed and the supervisor enters state 3.

State 3) The color histogram is used to recover the face (histogram always returns a result). If the tracking CF is high again (>0.5), the correlation template is re-initialised at an eye position estimated from the face position and the correlation is run again (i.e., the system reverts to state 2). If, on the other hand, the tracking CF drops below a threshold (i.e., CF<0.5), the supervisor enters state 1.

6. Conclusion

The complementary nature of multiple perceptual techniques means that they can be combined to build a system which is robust and flexible. Active camera control and active processing control are necessary for real time performance. Focussing the attention of perceptual processes on small regions of interest permits the system to perform image processing in real time without loss of field of view. The SERVP architecture permits a number of such component processes to be combined and coordinated. The system supervisor coordinates the execution of processes according to their success and to the success of the overall system. The SERVP architecture permits the supervisor to interpret the user's movements by changing the processes which are activated. In this manner the user may accomplish his task without distractions of remaining in a narrow field of

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996

view, or giving commands to change processing modes.

The cooperation of multiple visual processes is primarily intended for improving the robustness of the tracking system. It can also serve as the foundation for increasing the transmission bandwidth. Low bandwidth is an important source of discontinuities within the human communication process [10]. It can be improved by using high speed ATM based-technology, complemented with compression techniques. In CoMedi, we obtain very high video compression ratio by stabilising the image of the user using our cooperating techniques. Stabilisation provides the position, scale and illumination normalisation which makes possible compression using an incremental orthogonal space. This compression process is made possible by robust and precise face tracking.

Acknowledgement

This work is supported by France Telecom-CNET. We are grateful to F. Kirouche and F. Planchon for the design and implementation of the face tracking processes, and to Joachim Leber and Frederick Berthommier for collaboration in developing the processes for estimation of position from speech utterances.

References

[1] J. L. Crowley, P. Stelmaszyk, T. Skordas and P. Puget, "Measurement and Integration of 3-D Structures By Tracking Edge Lines", International Journal of Computer Vision, Vol 8, No. 2, July 1992.

[2] J. L. Crowley and Y. Demazeau, "Principles and Techniques for Sensor Data Fusion", Signal Processing, Vol 32 Nos 1-2, p5-27, May 1993

[3] J. L. Crowley and H. I Christensen, Vision as Process, Springer Verlag, Heidelberg, 1994.

[4] J. L. Crowley and J. Martin, "Experimental Comparison of Correlation Techniques", IAS-4, International Conference on Intelligent Autonomous Systems, Karlsruhe, March 1995.

[5] J. L. Crowley and J. M. Bedrune, "Integration and Control of Reactive Visual Processes", 1994 European Conference on Computer Vision, (ECCV-'94), Stockholm, may 94.

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

[7] W. Gaver, G. Smets, and K. Overbeeke, "A Virtual Window on a Media Space", Proc. CHI'95, ACM publ., pp. 257-264, 1995.

[8] R. Kraut, C. Cool, R. Rice, and R. Fish, "Life and Death of New Technology: Task, Utility and Social Influences on the Use of a Communication Medium", Proc. CSCW'94, ACM Publ., pp.13-21, 1994.

[9] Scheile B. and Weibel, A., "Gaze Tracking Based on Face Color", International Workshop on Face and Gesture Recognition, Zurich. July 1995.

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

[11] M. J. Swain and D.H. Ballard,Color Indexing, IJCV, Vol 7, No 1, p11-32, 1991.

	Correlation	**Eye blink**	**Color Histogram**
Precision (for a 100X100 image)	1 pixel	2 pixels	4 pixels
Stability (resistanc to noise)	+++	++	+
Efficiency (cycle time for a 192 x 144 image)	24 images/s	16 images/s	12,5 images /s
Fiability (capacity to locate a face)	++	++	+++
Tolerance to light variations	++	+++	++
Availability (Result on demand)	+++	+	+++
Initialisation required	yes	No	yes

Figure 6. Comparative results for the cross-correlation tracker, the eye blink detector and the color histogram match as implemented on a SGI Indy.

Published in the proceedings of the 2nd international conference on
Automatic Face and Gesture Recognition, Killington, Vermont, Oct 1996