# Robust Computer Vision for Computer Mediated Communication

**François Bérard, Joëlle Coutaz**

CLIPS-IMAG
BP 53
38041 Grenoble Cedex 9, FRANCE
{francois.berard, joelle.coutaz}@imag.fr

**James L. Crowley**

GRAVIR-IMAG
INPG, 46 av. Félix Viallet
38031 Grenoble, FRANCE
jim.crowley@imag.fr

**ABSTRACT** In Computer Mediated Communication, static video cameras provide a restricted field of view of remote sites. The concept of virtual window has been introduced to alleviate this problem. Although this technique is interactionally attractive, early implementations have proved unsatisfactory because of lack of robustness. In this paper, we demonstrate that robustness and accuracy can be increased by combining multiple computer vision techniques such as eye blink detection, skin color histogram and cross correlation, that adapt to a variety of operating conditions.

**KEYWORDS** Computer Mediated Communication, Virtual window, Computer vision, face tracking.

## 1. INTRODUCTION

In Computer Mediated Communication such as video conferencing, static 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, static cameras imply extra articulatory tasks that interfere with real world activity. In particular, users must keep their head (or an object of interest) within the field of the camera in order to be perceived by distant parties. The concept of virtual window tends to alleviate this problem by slaving a remote controlled camera to movements of the user's head (Gaver, 1995): when the head moves to the left, the camera rotates to the right, and vice versa. The same control is used for vertical head translations. The resulting behavior is similar to that which people get when looking through a real window.

The implementations of the Virtual Window concept by Gaver (1995) and Cooperstock (1995) use a simplistic vision technique based on background subtraction. Although this technique is computationally inexpensive, it exhibits three weaknesses that impede usability: first, an initialization step is required to register the background image. For so doing, the user has to move out of the field of view of the camera. In addition, the technique tracks any movement. As a result, it is not reliable for varying backgrounds. Third, the technique is sensitive to camera noise, provoking jittering of the head position estimation and consequently an inaccurate remote camera control. Gaver admits that his system was not accepted by users due primarily to the lack of robustness of the tracking system.

In this paper we demonstrate that robustness and accuracy can be increased by combining multiple computer vision techniques such as eye blink detection, skin color histogram and cross correlation, that adapt to a variety of operating conditions.

## 2. THREE TECHNIQUES FOR LOCATING HEADS

A human must periodically blink to keep his eyes moist. The fact that both eyes blink together provides a redundancy which permits blinking to be discriminated from other motions in the scene. 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 can handle a wide range of lightning conditions. As a result, it doesn't need calibration. In addition, it is computationally inexpensive: the face can be searched in the whole image in a wide range of scales. Unfortunately, eye blink occurs very irregularly and not very often (about once every 30 seconds): we use it to boostrap the two other detection techniques.

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. However, color detection shares the same limitation as background subtraction: since the pixels are not processed as a coherent set but as individuals, the technique is sensitive to camera noise resulting in jittering. To achieve accuracy, we use correlation tracking.

Cross-correlation operates by comparing a reference template (e.g., piece of the nose) to an image neighborhood at each position within a search region. Cross correlation is very accurate (i.e., 1 pixel): it provides the stability and accuracy needed for mapping fine grained movements of the user's head. It suffers however from two drawbacks: it is unable to track non-rigid motions such as a head rotation; and it is very expensive when it comes to comparing the target to the image in every position of the search region. We keep a high processing rate by reducing both the target size (we typically use 8x8 targets) and the search region. The counterpart is that the tracking often loses the head, either because the user has rotated his head, or moved abruptly. This problem is solved using the cooperation of the two other complementary detection techniques.

## 3. COOPERATION

In order to support cooperation, the output provided by each tracker is normalized and formalized: each tracker returns its estimation of the head position, a precision and a confidence factor. The fusion of the results is performed using a Kalman filter (Crowley, 1997). Figure 1 illustrates the cooperation of the three face trackers implemented as a Tcl supervisor. 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 (< 0.5) 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 (>0.5), 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 (i.e., CF<0.5), the supervisor draws upon the eye blink detector.

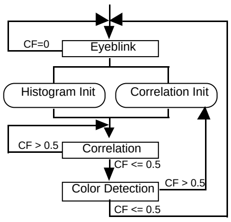

**Figure 1:** Cooperation of the three techniques.

## 4. CONCLUSION AND ACKNOWLEDGEMENTS

We have developed an autonomous, fast, accurate and robust face tracking system using the cooperation of multiple computer vision techniques. This work opens the way to the effective use of cameras as new input devices for interaction.

This work is supported by CNET France Telecom. We are grateful to F. Kirouche and F. Planchon for their participation in the implementation of the tracking system, and to N. Oliver for her insights on the color model.

## 5. REFERENCES

Cooperstock, J.R. Tanikoshi, K. and Buxton W. (1995) Turning Your Video Monitor into a Virtual Window. Proc. of *IEEE PACRIM, Visualisation and Signal Processing.*

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

Gaver, W. Smets, G. Overbeeke K. (1995) A Virtual Window on a Mediaspace. in Proc. *CHI'95*, ACM, Publ., pp. 257-264.