CapRe: a gaze tracking system in man-machine interaction

Christophe Collet (1,2), Alain Finkel(1), Rachid Gherbi(2)

(1)LSV-CNRS & ENS Cachan
61, av du Pdt Wilson
F-94235 Cachan Cedex, France
Tel : +33 1 47 40 24 04
Fax : +33 1 47 40 24 64

(2)LIMSI-CNRS
Bât. 508, BP 133
F-91403 Orsay, France
Tel : +33 1 69 85 81 21
Fax : +33 1 69 85 80 88

Email : collet@limsi.fr, gherbi@limsi.fr, finkel@lsv.ens-cachan.fr

Abstract — We present a real-time camera-based system designed for gaze tracking focused on human-computer communication. We aim to equip computer systems with a tool which can provide visual information about the user. This tool must satisfy interaction constraints and be not intrusive. Therefore we use a CCD camera, placed between the keyboard and the screen. The system detects the user presence, locates and then tracks his face, nose and both eyes. The detection is performed by combining image processing techniques and pattern recognition methods.

I. INTRODUCTION

Man-Machine Communication aims at enhancing the dialog between humans and machines. Recent work in spoken and written language processing, multimodal [1] and CSCW applications [2], and in new media integration contribute to a more natural human-machine interaction. However, the introduction of such new media must neither increase the cognitive load of the user nor disturb him. Indeed, to be more efficient, the user is supposed to focus on his main task. In this context, the exploitation of the user and environment visual information can truly enhance the interaction, provided that we use unobtrusive capture tools enabling free user movements.

In this context we are developing a capture system (CapRe) which locates and tracks the human face and its components in order to determine the gaze of the user, who is facing a workstation. The computer, by way of CapRe, "sees" if the user interacts with it and which area of the screen he looks at. Some ambiguous spoken and multimodal commands can then be avoided. CapRe can also be used a measuring tool for ergonomical and cognitive experiments using computers and/or for the evaluation of interfacing platforms.

CapRe proceeds in several phases. For each phases, the overall strategy is to perform either a simple computation on large image areas or a complex one on small regions. Thus, this deals with real-time processing constraint needed by both interfacing applications and cognitive experiments. The first three phases perform the localization and tracking of the face, nose and eyes. These phases alternate between two states: an initialization state and an adaptation state. In the initialization one, CapRe takes care of locating and recognizing the face components. In the adaptation state, it tracks these components in predictable closed neighborhood zones using the last established position. This reduces the search space and thus the computation time. Besides, CapRe updates its recognition parameters to take into account the eventual photometric and geometric variations of the components patterns. Besides, CapRe evaluates the orientation of the face (relatively to the screen) using a 3D projection model of the triangle represented by the nose and two eyes. Finally, it uses this orientation and the vector composed by the barycentres of the white and dark zones of eyes in order to obtain the gaze direction vector. The system goes on with these processes until it looses one of the face components. In this case, it leaves the adaptation state and returns to the initialization one.

The three first steps are implemented for both initialization and adaptation states. CapRe is able to detect the presence of a user, it locates his face components and it tracks them while the user is in the camera vision field. CapRe works at video frame rate (12 images per second) using a SGI Extreme workstation equipped with Galileo digitizer. We are now working on the quantitative evaluation of the system by recording a video sequence and the corresponding CapRe results. Before presenting some results, we describe here below the CapRe specification and functioning.

II. SPECIFICATIONS OF THE SYSTEM

We can give the following specifications for such gaze capture system involving the following constraints : the system must allow us to enhance the

Published in Proceedings of IEEE International Conference on Intelligent Engineering Systems INES'97, Septembre 15-17, 1997, Budapest, Hungary, pp.557-581.

human-computer interaction without disturbing anywhere the user. Hence, a non intrusive and quite silence tool is necessary. The user is facing the workstation and must be free in his movements. The location and the orientation of the camera should allow the acquiring of an exploitable images of the user and especially the image data representing his eyes. Besides, the system should be costless and must deliver a results in real time.

In CapRe project, we propose a system allowing to measure the gaze of a user facing the computer screen and working on any task. These measurements are computed from black and white video sequence data.

III. EXISTING SYSTEMS

We note that other existing systems have the same objective of gaze capture. We complete here below the bibliographic list established by Charbonnier [3]. We should distinguish between two main families : intrusive systems and non-intrusive systems, and within the last, those which are camera-based.

A. Intrusive systems

Some systems constraint the user to wear one or several captors which handicap his spontaneous gestures : a magnetic spark-coil included in a lens posed on the eye; some electrodes placed around the eye in order to measure the muscular activity; a glasses strongly attached on the head and equipped with several transmitters and infra-red sensors; a head device containing a semi-reflecting mirror located in face of the user eye and a micro-camera which captures the eye image reflected by this mirror.

Other systems are noisy or have some of their components in movement like for instance a semi-reflected mirror facing the screen, coupled with a rotation mechanism allowing us to track the movements of the user, in a manner to keep a fixed image reflected by the mirror and captured by the camera.

B. Non intrusive systems

It concerns here the capture systems based on active camera(s) placed somewhere on the desk. We can list systems using an infra-red cameras involving a light source of this kind. These systems can be perturbed by the infra-red rays coming from the environment lights but they allow a precise results less than one degree and with a high frequencies until 60 Hz. But, these systems being very expensive, especially when using the infra-red cameras, so the experimentations are done with a monochrome or color video camera (visible light).

C. Camera-based systems

Baluja et Pomerleau [4] used a color camera and a method based on a neural network for the detection and tracking of the eyes and also for the determination of the gaze. The system works at 15 Hz rate with a precision of one degree, but constraints the user to stay in place authorizing only very small head movements. On the other hand, the eyes are detected regarding the specular reflect of a light coming from a facing source, and hence obstruct the user, which is no acceptable for us.

Stiefelhagen, Yang et Waibel [5] designed a system which is close to our system in several points. It allows to detect and track the face, eyes, nose and mouth of a user using a color camera and without a specific light sources. It seems hence robust relatively to lighting variations. However, the user must stay for a long time in face of the camera in order to allow the system to detect the face and its components. Besides, the system use only the geometric constraints between these components in order to recognize them, but does not perform a pattern recognition processing to be sure that it detects the right components. Hence, if it confuses the components each together or with other element of the face (mustache, glasses, etc.), it has no mean to take into account these errors.

We propose a costless and real time system using a specific lighting, which perform the detection and tracking of the face, nostrils and eyes with a reliable results which avoids to constraint physically the user.

IV. EXPERIMENTAL PLATFORM

In order to satisfy the above specifications about the capture system, we have chosen to use a black and white camera which is silent (it is not associated with any mechanical systems : electric zoom, automatic control of the focus, etc.).

We made several tests on the location and the orientation of the camera. When comparing the grabbed images from the various points of view (on the screen, at side of the monitor and between the monitor and the keyboard), we stand that the last point of view is the best one because it allows us to discriminate all the orientations of the gaze. In consequence, we choose to use a small camera which could be positioned under the screen and behind the keyboard "Fig. 1". Besides, the camera orientation (being in counter-plunge) allows us to obtain images containing the ceiling as background. The ceiling objects are stable, so this should simplify the separation of the searched target (the user) from the image background.

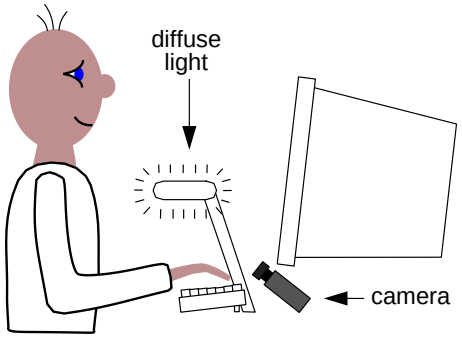

Fig. 1 : CapRe experimental platform.

A diffuse light source is placed on each side of the monitor in order to create a regular lighting without in the same time disturbing the user. These sources are oriented in a manner to make as less reflects as possible on the eyes which could be a source of difficulties during the image processing.

V. FUNCTIONING OF CapRe SYSTEM

Before obtaining the gaze, the CapRe system passes through several phases "Fig. 2". We recall here that the application context (human-computer interaction) requires to design a real-time system. Our global approach is to either perform a simple (fast) computations on a large regions of the image or a complex computations (long) on a small areas. Hence, each phase uses one or both of these two processing strategies.

Except the phases 4 and 5 which could be performed in parallel, each phase delivers a results which should be used by the following phase. The first three phases realize the localization and the tracking of the face, nose and the eyes. These phases alternate between two states : an initialization state and an adaptation one. The initialization state consists on locating the face and its components (nose and eyes). The associated processing can take a sufficient time in order to be sure that it detects the right locations. The adaptation state focuses on tracking the face and its components. This tracking is done by searching each component in the neighbor regions of its previous location (previous image). Indeed, this reduces the search space, and the computation time in consequence. Besides, in this same state, the system adapts the recognition parameters of each component according to the currant image.

First of all, the system is in the initialization state. When it considers a significant user movement, it tries to locate a face and its components. If it finds the face from several consecutive images, the system passes in the adaptation state. It should stay in this state until it considers that it lost the face. In this case, it backs in the initialization state. This alternation is transparent for the user and it allows us to avoid doing a heavy calibration step each time the system lost the face. The phases 4, 5, and 6 are performed only when the system is in the adaptation state.

1st phase Face

This phase consist in determining and making the extraction of the bounding box of the user face. This is done by computing the difference of two consecutive images in order to detect the moving parts. This treatment is performed on a sub-sampling of the image because, on one hand the object to be detected is large relatively to the image dimensions and on the other hand the number of pixels to be processed is strongly reduced. In practice, the system takes into account only one line every five lines in each image. If the system detects very little movements, it means that the user doesn't move and so it use the bounding box detected in the previous images. This method enable the system to reject absurd bounding box detected in this case.

2nd and 3rd phases Nose and Eyes

In each of the following phases (2 and 3), we apply two successive processes in order to respectively detect the nose and eyes. Firstly, we detect the dark zones which could be considered as nostril or pupil. To obtain these zones, our process is based on a thresholding operation followed by a pixels fusion one. Those zones which are not physiologically consistent according to size and proportions of nostril or pupil, are rejected. However, this process generates several potential zones (more than those representing the nose and eyes). It is hence necessary to apply a second process on these zones in order to determine which of them correspond to the searched targets. Considering each zone, this second process consists in correlating two gradient profiles : the target gradient profile (nose or eye) with the zone gradient one. For the robustness of the process, the gradient profile take into account several pixels along an horizontal line. The use of two processes respects the two strategies of computation described above (i.e. simple operations of pixels fusion on large regions followed by a cost correlation, but applied only on a few small zones).

Published in Proceedings of IEEE International Conference on Intelligent Engineering Systems INES'97, Septembre 15-17, 1997, Budapest, Hungary, pp.557-581.

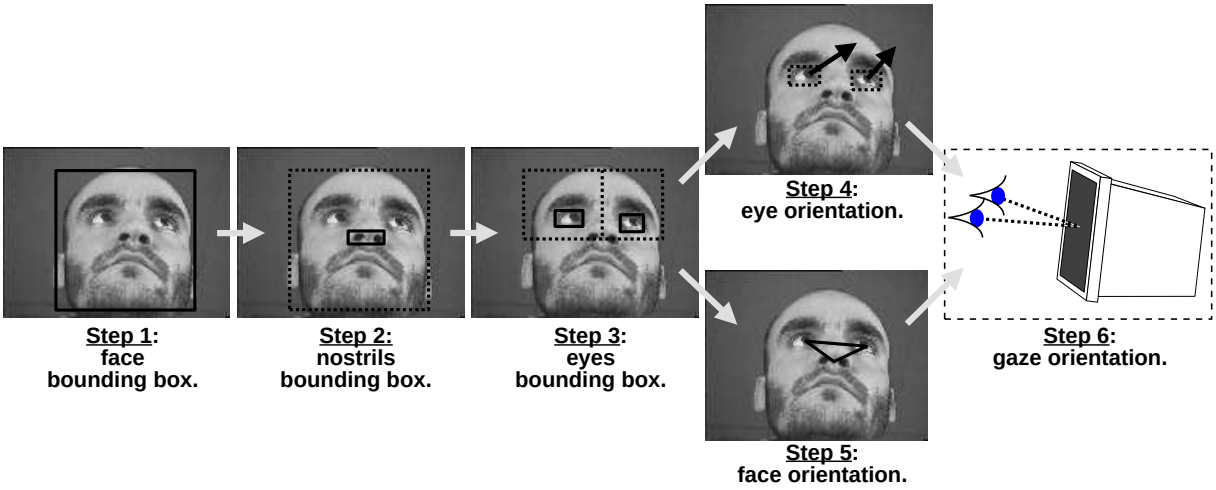

Fig. 2: Functional diagram of CapRe system.

2^{nd} phase Nose

The algorithm developed in this phase takes into account only pixels which are in the face bounding box detected in the phase 1. It concerns the determination and extraction of the nostrils bounding box. We apply here the two processes described above involving the dark zones detection and the correlation of the nostril gradient profile. We decide to firstly detect the nostrils before the eyes because they are geometrically and photometrically more stable. Besides, the nose is rarely masked (by glasses for instance) and its inter-persons variability is very weak. Finally, considering the location of the nose, it is more easy to suppose that the eyes could be searched on the top of this location, one on each side.

3^{rd} phase Eyes

The algorithm developed in this phase takes into account only pixels which are in the left-top and right-top of the nose bounding box detected in the phase 2. It aims at the extraction of the two eyes bounding boxes. It works likewise as for the nostrils algorithm but using here the eyes (pupil) gradient profile.

At the end of these three phases, we obtain all the bounding boxes of the face, nostrils and eyes for each image of the video sequence. We can now try to determine the gaze. This could be done by the algorithms described in the following phases.

4^{th} phase Eye orientation

In this phase the system compute an evaluation of the orientation of the gaze relatively to the user face using the pixels of the eyes bounding boxes. This orientation is obtained by the vector composed by the two following points: the center of the pupil and the barycentre of the eye white surface. The center of the pupil coincides with the darkest pixel. The white surface of the eye is extracted using a thresholding operation (the threshold is determined from the local histogram of the eye bounding box). Indeed, each eye bounding box could give a different threshold.

5^{th} phase Face orientation

Here the evaluation of the orientation of the user face relatively to the screen is done. It use the four points representing the four extremities of the eyes and the base point of the nose. This is done by 3D modeling of the projection of the triangle represented by the previous points D.M. Gavrila et L.S. Davis [6].

6^{th} phase Gaze orientation

The last phase makes the evaluation of the gaze of the user relatively to the screen by composing the two vectors determined in the two previous phases.

VI. RESULTS

All the components of the system are developed on SGI–INDIGO workstation equipped with a "Galileo" video grabber. This workstation is powerfully sufficient for all the developed processes (10 ms for the processing of a 384x144 image in the worst case - initialization state). However the "Galileo" video grabber limits the capture frequency proportionally to the image dimensions. The CapRe system can detect the user face and its components, it then tracks them at 12 Hz image frequency for a video sequence of 384x144 images. But the the theorical frequency could be greater than the 25 Hz PAL standard one for a 768x288 images. Hence, our system can work at higher frequency than camera-based systems listed in the introduction. Besides, we gathered a corpus of 40 video films representing 32 different persons facing the screen. The duration of each film is approximately 1 minute. We have done a first evaluation

on this corpus for the phases 1 and 2 in initialization state. We obtain 3.4% of errors for face bounding box detection, 11.5% of errors for nose localisation and 79% of good detection and localisation (less than 5 mm) of both nostrils.

VII. CONCLUSION

Except the phase 6, all other phases were implemented for the two states (initialization and adaptation). The phases 1, 2 and 3 are integrated in the CapRe system. After the evaluation of these phases, we should integrate the rest of phases. CapRe is able to detect the presence of the user, to locate his face, nose and eyes, and to track them in real-time when they are visible by the camera. We are working on the quantitative evaluation of the system. This evaluation is very important for the future of the system and also for an eventual comparison with other systems.

We plan to integrate the CapRe system into applications of human-computer interfaces on one hand, like the multimodal projects (Meditor of Bellik [7], Mix3D of Bourdot, Krus et Gherbi [8]) and the LSF gesture recognition system of Braffort [9]. On the other hand, we envision to use CapRe as a gaze measurement tool in a cognitive experimentations which aim at analyzing the various strategies used by a user during tasks involving software training or Web information searching.

VIII. ACKNOWLEDGMENT

Thanks to the French Ministry of High-Education and Research for giving research grant to C. Collet PhD. The authors thank B. Doval and A. Braffort for their valuable comments.

IX. REFERENCES

[1] Minh Tue Vo, Ricky Houghton, Jie Yang, Udo Bub, Uwe Meier, Alex Waibel, and Paul Duchnowski. Multiodal Learnig Interfaces. In *Proc. of ARPA, Spoken Language Technology Workshop*, Barton Creeks, January 1995.

[2] Hiroshi Ishii, Minoru Kobayashi, and Kazuho Arita. Iterative Design of Seamless Collaboration Media. *Communication of the ACM*, 37(8):pp. 83–97, August 1994.

[3] Colette Charbonnier. *La commande oculaire: étude et validation exprérimentale d'interfaces homme-machine contrôlées par la direction du regard.* PhD thesis, Université Joseph Fourier / LETI-CEA, Grenoble, France, October 1995.

[4] Shumeet Baluja and Dean Pomerleau. Non-Intrusive Gaze Traking Using Artificial Neural Networks. Technical report, ISL, Carnegie Mellon University, Pittsburgh, PA 15213, 1994.

[5] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. A Model-Based Gaze Tracking System. In *Proc. of IEEE Intl. Joint Symposia on Intelligence & Systems - Image, Speech & Natural Language Systems*, Washington DC, USA, 1996.

[6] Dariu M. Gavrila and Larry S. Davis. 3-D Model-based Tracking of Humans in Action: a Multi-view proach. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, June 1996.

[7] Yacine Bellik, Stéphane Ferrari, Françoise Néel, and Daniel Teil. Interaction Multimodale : Concepts et Architecture. In *Actes d'Interface des Mondes Réels et Virtuels*, Montpellier, France, June 1995.

[8] Patrick Bourdot, Mike Krus, and Rachid Gherbi. MIX 3D : une plate-forme expérimentale pour des interfaces multimodales dédiées à la CAO. In *IHM'95, Septièmes Journées sur l'Ingénierie des Interfaces Homme-Machine*, 1995.

[9] Annelies Braffort. A gesture recognition architecture for sign language. In *Proc. of ASSETS'96, ACM*, Vancouver, Canada, November 1996.