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 non intrusive. Therefore we use a CCD camera, placed between the keyboard and the screen. The system detects the user's presence, locates and then tracks his face, nose and both eyes. The detection is performed by combining image processing techniques and pattern recognition methods.

1 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 visual information concerning the user and the environment 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, recognizes that the user is interacting with it and which area of the screen he is looking at. Some ambiguous spoken and multimodal commands can then be avoided. CapRe can also be used as a measuring tool for ergonomical and cognitive experiments using computers or for the evaluation of interfacing platforms.

CapRe proceeds in several phases. For each phase, 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 constraints 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 step, CapRe takes care of locating and recognizing the face's 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. Additionally, CapRe updates its recognition parameters to take into account the eventual photometric and geometric variations of the components patterns. Further, 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. During its functioning, CapRe toggles between these two states.

The first three 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 tracks them while the user is in the camera field of vue. CapRe works at video

Published in International Journal of Advanced Computational Intelligence, ACI, vol.2, n°3, pp.77–81, L. A. Zadeh, T. Fukuda & K. Hirota Eds, 1998.

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 the CapRe specification and functioning.

2 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 human-computer interaction without disturbing the user in any way. Hence, a non intrusive and noiseless tool is necessary. The user is facing the workstation and must be free to move. The location and the orientation of the camera should allow the acquisition of exploitable images of the user and especially the image data representing the eyes. Additionally, the system should be cheap and deliver a results in real time.

In the CapRe system, we propose a tool 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.

3 Existing systems

We note that other existing systems have the same objective of gaze capture. We complete here below the bibliographical 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.

3.1 Intrusive systems

Some systems constrain 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. Most of these tools provide accurate measurements but are not suitable for natural interaction.

Some 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 order to keep a fixed image reflected by the mirror and captured by the camera.

3.2 Non intrusive systems

Such systems are based on active camera(s) placed somewhere on the desk. Some use 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 precise results (less than one degree) and with a high frequencies (up to 60 Hz). But, they are very expensive, especially when using the infra-red cameras, so the experimentations are done with a monochrome or color video camera (visible light).

3.3 Camera-based systems

Baluja and 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 constrains 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 light coming from a facing source, and hence obstructs the user, which is not acceptable for us.

Stiefelhagen, Yang and Waibel [5] designed a system which is close to our system in several points. It allows the detection and tracking of the face, eyes, nose and mouth of a user using a color camera and without specific light sources. It seems hence robust relatively to lighting variations. However, the user must stay for a long time in front of the camera in order to allow the system to detect the face and its components. Besides, the system uses 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. In consequence, if it confuses the components with other element of the face (mustache, glasses, etc.), it has no means takeing into account these errors.

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

4 Experimental platform

In order to satisfy the above specifications regarding 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 top of the screen, on one side of the monitor and between the monitor and the keyboard), we noted 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 (see Figure 1). Furthermore, the camera's 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 background.

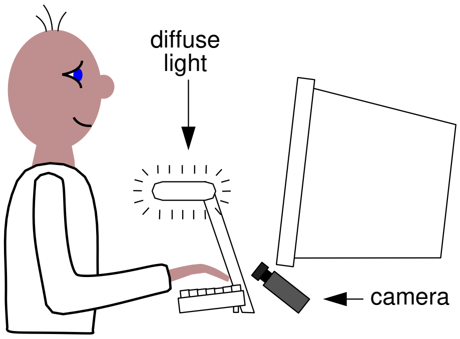

Fig. 1 : CapRe experimental platform.

A diffuse light source is placed on each side of the monitor in order to create a regular lighting without disturbing the user. These sources are oriented in such a way as to minimize reflection on the eyes (which could be a source of difficulties during the image processing) as much as possible.

5 Functioning of CapRe

Before obtaining the gaze direction, the CapRe system passes through several phases (see Figure 2). The application context (human-computer interaction) requires the design of a real-time system. Our global strategy consists on performing either a simple (fast) computation on a large region of the image or a complex computation (long) on a small area. Indeed, each phase uses one or a combination of these two processing strategies.

Except 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 eyes. These phases alternate between two states : an initialization state and an adaptation one. The initialization state consists in locating the face and its components (nose and eyes). The associated processing can take 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). This reduces the search space and the computation time. Furthermore, in this state, the system adapts the recognition parameters of each component according to the current 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 has lost the face. In this case, it returns to the initialization state. This switching is transparent for the user and it allows us to avoid doing a heavy calibration step each time the system loses the face. The phases 4, 5, and 6 are performed only when the system is in the adaptation state.

1st phase Face

This phase consists in determining and extracting of the bounding box of the user's face. This is done by computing the difference of two consecutive images in order to detect the moving parts. This processing is performed on a sub-sampling of the image because, on one hand the object to be detected is large relatively to the size of the image and on the other hand the number of pixels to be processed is greatly reduced. In practice, the system takes into account only one in five lines of each image. If the system detects very little movements (which means that the user didn't move), the bounding box detected in the previous image should be used. This method enables 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 detect respectively 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 pixel fusion. The 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. For the robustness of the process, the gradient profile takes 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).

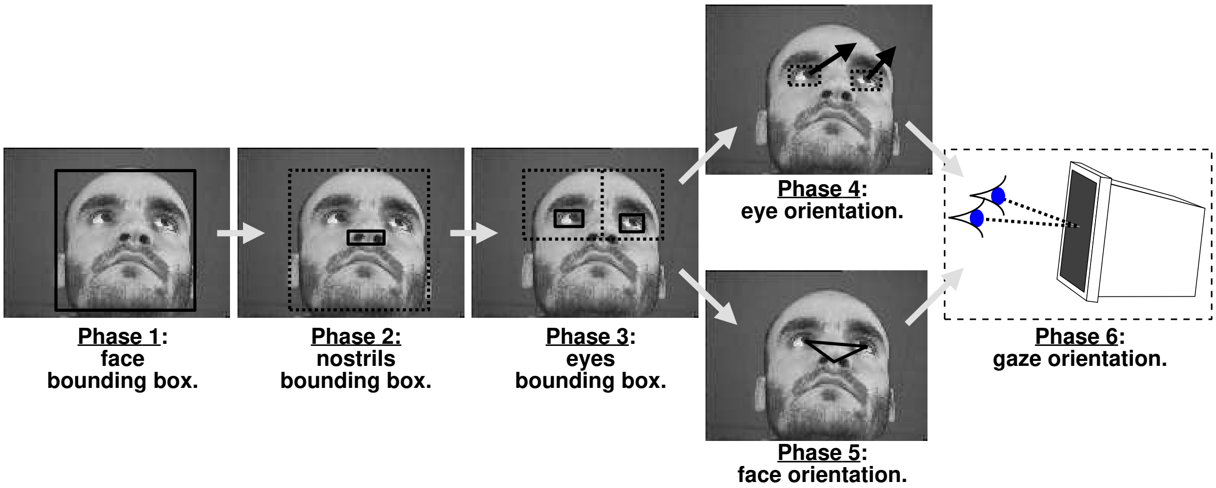

Fig. 2: Functional diagram of CapRe system.

2nd 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 extracts the nostrils bounding box. We apply here the two processes described above involving the dark zones detection and the correlation of the nostrils' gradient profile. We decided to 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-person variability is very weak. Finally, considering the location of the nose, it is easier to suppose that the eyes could be searched above this location, one on each side.

3rd phase Eyes

The algorithm developed in this phase takes into account only pixels which are in the top-left and top-right of the nose's bounding box detected in the phase 2. It aims at the extraction of the two eyes' bounding boxes. It works like 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.

4th phase Eye orientation

In this phase the system computes an evaluation of the orientation of the gaze relatively to the users face using the pixels in 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's white area. The center of the pupil coincides with the darkest pixel. The white area of the eye is extracted using a thresholding operation (the threshold is determined from the local histogram of the eye's bounding box). Indeed, each eye's bounding box could give a different threshold.

5th phase Face orientation

In parallel with the phase 4, the system evaluates the user's face orientation relatively to the screen. It uses 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 and L.S. Davis [6]).

6th phase Gaze orientation

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

6 Results

All the components of the system were developed on SGI–INDIGO workstation equipped with a "Galileo" video grabber. This workstation is 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 images' size. The CapRe system can detect the user's face and its components, it then tracks them at 12 Hz image frequency for a video sequence of 384x144 images. But the theoretical frequency could be greater than the 25 Hz PAL standard for 768x288 images. Hence, our system can work at higher frequency than camera-based systems listed in the introduction.

Additionally, we gathered a corpus of 40 video films (15000 images) 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 3 in initialization and the phase 2 in initialization and adaptation state. We consider that there is an error for the face's bounding box localization if it doesn't include the nose and both eyes. We consider that the nose or the eyes localization is an error if the distance between it and the real localization is greater than 5 millimeters.

state	Phase 1 face bound- -ing box	Phase 2 nose's center	Phase 3 pupils' center
init.	1.1%	12.7%	24.5%
adapt.		3.8%	

Fig. 3 : Results diagram for the first 3 phases.

7 Conclusion

Except the 6th phase, 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 will integrate the other 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 and 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 experimentation which aims at analyzing the various strategies used by a user during tasks involving software training or Web information searching.

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.

References

[1] Minh Tue Vo, Ricky Houghton, Jie Yang, Udo Bub, Uwe Meier, Alex Waibel, and Paul Duchnowski. Multimodal Learning 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.* Mémoire de doctorat, 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.