CapRe : un système de capture du regard dans un contexte d'interaction homme-machine

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

(1)LSV-CNRS & ENS de Cachan
61, av du Pdt Wilson
F-94235 Cachan Cedex, France
Tél : +33 01 47 40 24 04
Fax : +33 01 47 40 24 64

(2)LIMSI-CNRS
Bât. 508, BP 133
F-91403 Orsay, France
Tél : +33 01 69 85 81 21
Fax : +33 01 69 85 80 88

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

Résumé

Nous présentons dans ce papier un système de vision temps-réel pour la capture du regard dans le cadre de la communication homme-machine. Notre objectif est de doter un système informatique d'un outil qui exploite les informations visuelles concernant l'utilisateur. Cet outil doit respecter les contraintes liées à l'interaction. En particulier, il ne doit pas être intrusif ni gêner l'utilisateur dans ses mouvements naturels. Nous avons choisi d'utiliser une caméra noir et blanc, placée entre le clavier et l'écran. Le système détecte la présence d'un utilisateur. Il localise ensuite son visage, son nez et ses deux yeux, et il les suit dans la séquence d'images. Le système passe alternativement par deux états : initialisation et adaptation. La détection est réalisée en combinant des techniques de traitement d'images avec des méthodes de reconnaissance de formes.

Mots clefs : suivi du regard, vision, reconnaissance de formes, communication homme-machine.

Abstract

We present in this paper a real-time vision system designed for gaze tracking focused on human-machine communication. We aim to equip computer systems with a tool which can provide visual information about the user. This tool must satisfy interaction constraints. To be more specific, it should be neither intrusive nor restrictive in order to allow natural interactions. Therefore we use a black and white CCD camera, placed between the keyboard and the screen. First of all, the system detects the user's presence and it locates the face, the nose and both eyes. Then, these features are tracked through the image sequence. The system toggles between two states : initialisation and adaptation. The detection is performed by combining computer vision technics and pattern recognition methods.

Keywords : gaze tracking, computer vision, pattern recognition, human-machine communication.

Catégorie : outils, présentation orale.

Publié dans les actes des 6ème Journées Internationales sur l'interface des mondes réels et virtuels, 28-30 mai 1997, Montpellier, France.

1. Introduction

La Communication Homme-Machine (CHM) s'intéresse à l'amélioration des échanges d'informations entre les hommes et les machines. En particulier, les travaux récents en multimodalité (en entrée et en sortie), en travail coopératif (collecticiel), en reconnaissance du langage parlé et écrit, et dans l'intégration de nouveaux médias, permettent un dialogue homme-machine plus naturel [Cad93]. Cette amélioration des Interfaces Homme-Machine nécessite l'introduction de nouveaux canaux de communication, sans pour autant augmenter la charge cognitive de l'utilisateur. En effet, l'interface doit lui permettre de se focaliser sur la tâche qu'il souhaite accomplir pour être plus efficace [FA93]. La prise en compte des gestes de l'utilisateur contribue à un dialogue naturel, à condition de disposer d'outils de capture discrets pour que l'utilisateur puisse effectuer des mouvements spontanés. Ainsi l'exploitation des informations visuelles sur l'utilisateur, notamment ses mouvements oculaires, permet d'améliorer la communication de diverses manières.
Un ordinateur capable d'enregistrer et de traiter automatiquement des informations visuelles et celles concernant l'interaction, est un outil de mesure pour des expériences ergonomiques ou cognitives dans le cadre de l'utilisation d'une machine informatique, et pour l'évaluation de plate-formes matérielles et/ou logicielles dans le cadre des interfaces Homme-Machine [ABA92].
L'ordinateur "voit" si l'utilisateur veut interagir avec lui. Cela permet d'intégrer une nouvelle modalité dans des systèmes d'interfaces multimodales, par exemple pour résoudre certaines ambiguïtés liées à la commande vocale : "l'utilisateur s'adresse t-il à la machine ou à un autre interlocuteur ?".
L'ordinateur "voit" dans quelle zone de l'écran l'utilisateur regarde. Dans le cadre d'une interface multimodale cela permet de contextualiser des commandes (vocales ou tapées au clavier) selon la fenêtre ou l'objet sur lequel interagit l'utilisateur. De manière générale, cela peut servir d'outil de pointage alternatif à la souris, comme l'ont montré les études menées notamment par Charbonnier et Massé [CM94]. Cela peut aussi être intégré à un système de réalité virtuelle comme le suggère Jacob [Jac95]. Enfin, pour les logiciels de type didacticiel, il peut être utile de savoir si l'utilisateur regarde l'objet désigné à l'écran.
L'ordinateur "voit" s'il y a un utilisateur, ce qui permet par exemple, le déclenchement et l'arrêt de l'économiseur d'écran, ou encore la fermeture automatique du compte.

Nous pouvons donc donner les spécifications suivantes à un système de capture du regard pour qu'il satisfasse les contraintes suivantes : le système doit permettre d'améliorer la communication sans risquer de gêner l'utilisateur. Ainsi un outil de capture non-intrusif et silencieux est nécessaire. L'utilisateur est face à son poste de travail et doit être libre de ses mouvements. La position et l'orientation de la caméra doivent permettre la capture d'images exploitables de l'utilisateur et notamment de ses yeux. D'autre part, le système ne doit pas avoir un coût trop élevé et doit donner des résultats en temps réel.

Nous présentons un système qui permet de mesurer la direction du regard d'un utilisateur face à un écran d'ordinateur, à partir d'images captées par une caméra noir et blanc. Il existe déjà quelques systèmes de capture du regard, nous complétons la liste établie par Charbonnier [Cha95]. On peut distinguer deux familles :

– Les systèmes intrusifs : certains systèmes nécessitent de porter des capteurs qui handicapent les mouvements spontanés de l'utilisateur : une bobine magnétique intégrée dans une lentille posée sur l'œil; des électrodes placées autour de l'œil pour mesurer l'activité musculaire; des lunettes solidement attachées à la tête et équipées d'émetteurs et de capteurs infrarouge;

un casque contenant un miroir semi-réfléchissant placé devant l'œil de l'utilisateur et une micro-caméra qui capte l'image de l'œil reflétée par le miroir.

D'autres systèmes sont bruyants ou ont des parties en mouvements comme par exemple un miroir semi-réfléchissant placé devant l'écran, couplé à un système de rotation permettant de suivre les mouvements de l'utilisateur, de manière à garder une image fixe reflétée par le miroir dans une caméra.

– Les systèmes non-intrusifs : il s'agit de systèmes de capture par caméra placée sur le bureau. On trouve des systèmes utilisant des caméras sensibles à la lumière infrarouge, nécessitant une source lumineuse de ce type. Ces systèmes sont perturbés par les infrarouges provenant de la lumière du jour mais permettent de faire des mesures d'une précision inférieure au degré et avec une fréquence allant jusqu'à 60 Hz. Ces systèmes étant très onéreux, notamment à cause de la caméra infrarouge, des recherches ont été réalisées avec des caméras vidéo couleur et monochrome (lumière visible).

Baluja et Pomerleau [BP94] utilisent une caméra couleur et des réseaux de neurones pour la détection et le suivi des yeux, ainsi que l'évaluation de la direction du regard. Le système fonctionne à 15 Hz avec une précision d'un degré, mais oblige l'utilisateur à rester dans la même position, avec peu de mouvements de la tête. D'autre part, les yeux sont détectés grâce au reflet spéculaire d'un éclairage de face gênant pour l'utilisateur, ce qui est contraire à ce que nous souhaitons.

Stiefelhagen, Yang et Waibel [SYW96] ont mis au point un système dont le fonctionnement est proche du notre. Il permet de détecter et de suivre le visage, les yeux, le nez et la bouche d'un utilisateur avec une caméra couleur, et sans éclairage spécifique. Il est donc à même de s'adapter aux variations lumineuses. Mais l'utilisateur doit se tenir bien en face de la caméra pour que le système détecte le visage et ses composantes. De plus, Le système utilise leurs contraintes géométriques intrinsèques pour les reconnaître, mais n'applique pas de reconnaissance des formes pour être sûr d'avoir trouvé les bonnes composantes. Ainsi s'il confond les composantes entre elles ou avec un autre élément du visage, il n'a pas de moyen pour se rendre compte de son erreur.

Nous proposons un système peu onéreux qui avec l'appui d'un éclairage spécifique, permet de détecter le visage, le nez et les yeux de manière fiable et rapide, sans contrainte pour l'utilisateur.

2. Dispositif expérimental

Afin de satisfaire les spécifications du système de capture exposées dna l'introduction, nous avons choisi une caméra noir et blanc, silencieuse (elle n'est associée à aucun système mécanique : zoom électrique, réglage automatique de la mise au point ou système de recadrage).

Nous avons fait une série de tests sur la position de la caméra, en comparant les images captées depuis différents points de vue : au dessus du moniteur, à coté du moniteur et entre le moniteur et le clavier. Nous avons constaté que ce dernier point de vue permet de discriminer toutes les directions du regard. Nous avons donc choisi une caméra de taille assez petite pour être glissée entre le moniteur et le clavier [Fig.I]. Cet emplacement donne en outre un champ de vision en contre-plongée et permet donc d'avoir le plafond pour fond d'image. Le plafond étant stable, ceci simplifie la séparation entre le fond de l'image et ce que l'on veut détecter (i.e. l'utilisateur).

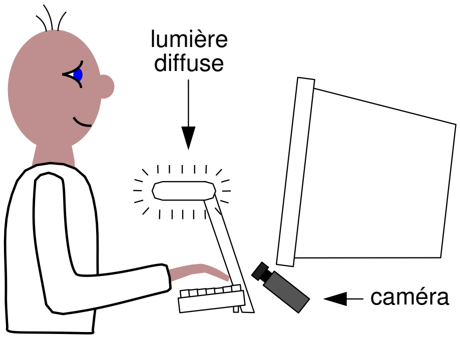

Fig.I : la plate-forme de CapRe.

Une source lumineuse diffuse, est placée de part et d'autre du moniteur de manière à créer un éclairage régulier sans pour autant risquer de gêner l'utilisateur, ni de provoquer des reflets de lumière dans les yeux gênants pour le traitement d'image.

3. Fonctionnement du système CapRe

Avant d'aboutir à la capture du regard, le système passe par plusieurs phases [Fig.II]. Rappelons que le contexte de l'interaction homme-machine exige un système en temps réel. Nous avons donc décidé de réaliser soit des calculs rapides dans une grande région de l'image, soit des calculs longs dans une petite région de l'image. Ainsi chacune des phases comporte l'une ou les deux stratégies de traitement.

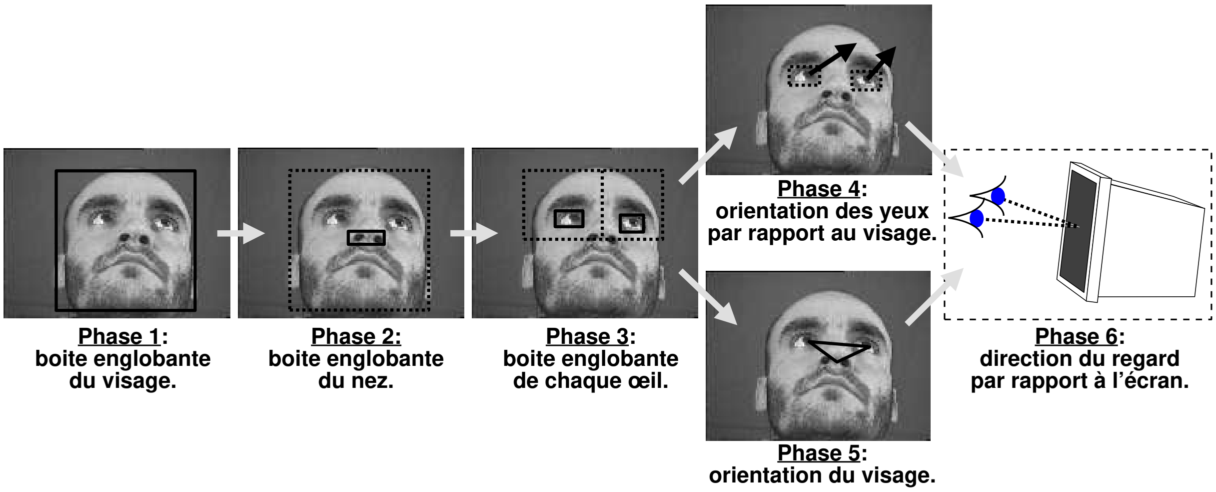

Fig.II : schéma fonctionnel du système CapRe.

Chaque phase donne un résultat qui est utilisé par la phase suivante, sauf pour les phases 4 et 5 qui pourrait être exécutées en parallèle. Les trois premières phases réalisent la localisation et le suivi du visage, du nez et des yeux. ces phases fonctionnent selon deux états : un état d'initialisation et un état d'adaptation. L'état d'initialisation consiste à localiser le visage et ses composantes (nez et yeux). Le traitement associé peut prendre du temps car Il est nécessaire de s'assurer que la localisation a bien été réalisée. L'état d'adaptation consiste à suivre le visage et ses composantes. Ce suivi se fait en recherchant chaque composante dans une région voisine de sa localisation dans l'image précédente. Cela réduit l'espace de recherche, donc le temps de calcul est très faible. De plus, le système adapte ses paramètres de reconnaissance pour chaque composante.

Dans un premier temps, le système est dans l'état d'initialisation. Dès qu'il détecte un mouvement, il tente de localiser un visage et ses composantes. S'il trouve ce visage sur plusieurs images successives, le système passe dans l'état d'adaptation. Il reste dans cet état jusqu'à ce qu'il perde le visage. Dans ce cas, il revient dans l'état d'initialisation. Les phases 3, 4 et 5 ne sont utilisées que quand le système est dans l'état d'adaptation.

1ère phase Visage : extraction de la boite englobante du visage de l'utilisateur en calculant la différence de deux images successives afin d'extraire les parties en mouvements.

2ème phase Nez : extraction de la boite englobante du nez à partir de la boite englobante du visage de l'utilisateur. On détecte dans un premier temps, les zones sombres susceptibles d'être les narines, puis on applique localement à ces zones un algorithme de reconnaissance par corrélation avec la forme du gradient du nez. Nous avons choisi de détecter en premier le nez, au lieu des yeux ou de la bouche, car il a une plus grande stabilité photométrique et géométrique, il est rarement occulté (par exemple par des lunettes) et car sa variabilité inter-personnes est très faible.

3ème phase Yeux : extraction de la boite englobante de chaque œil à partir de deux zones situées au dessus et de part et d'autre du nez. On détecte dans un premier temps, les zones sombres susceptibles d'être les pupilles, puis on applique localement à ces zones le même algorithme de reconnaissance qu'en phase 2, avec la forme du gradient de l'œil.

4ème phase Orientation du regard : évaluation de l'orientation du regard de l'utilisateur par rapport au visage à partir de la boite englobante de chaque œil, par calcul d'un vecteur composé par le centre de la pupille et le barycentre du blanc de l'œil.

5ème phase Orientation du visage : évaluation de l'orientation visage de l'utilisateur par rapport à l'écran, à partir de la position du nez et des yeux dans l'image par projection sur un modèle 3D comme le décrit l'article de D.M. Gavrila et L.S. Davis [GD96].

6ème phase Direction du regard : évaluation de la direction du regard de l'utilisateur par rapport à l'écran en composant les vecteurs calculés dans les deux phases précédentes.

4. Résultats

Nous utilisons une carte d'acquisition vidéo "Galileo" sur une station SGI–INDIGO II. Cette machine est suffisamment rapide pour le traitement d'image que nous faisons (10 ms pour le traitement d'une image 384x144, dans l'état d'initialisation). Cependant la carte d'acquisition limite la fréquence d'échantillonnage proportionnellement à la taille de l'image. Le système peut détecter un visage et ses composantes, et les suivre en temps réel, avec une fréquence de capture de 12 Hz, pour des images de 384x144, mais la fréquence théorique est de l'ordre de 25 Hz, pour des images de 768x288. Notre système peut donc fonctionner à des fréquences plus élevées que les autres systèmes de vision cités dans l'introduction.

5. Conclusion

Nous avons implémenté les trois premières phases du système de capture, pour les deux états, initialisation et adaptation. Le système est capable de détecter la présence d'un utilisateur, de localiser son visage, son nez et ses deux yeux, et de suivre ses composantes tant qu'elles sont dans le champ de vision de la caméra. Nous sommes en train de mettre au point une évaluation quantitative du système de capture pour la détection et le suivi des composantes du visage, en enregistrant les images captées et les résultats en sortie de chaque phase.

Il est envisagé d'intégrer notre système de capture du regard dans d'autres projets, comme les systèmes d'interfaces multi-modales : Meditor de Bellik [BFNT95], Mix3D de Bourdot, Krus et Gherbi [BKG95], et de Martin [Mar95], ou le système de reconnaissance de langue des signes de Braffort [Bra96].

Nous utiliserons notre système pour la mise au point d'expériences cognitives. Celle-ci visent à analyser les différentes stratégies mises en œuvre par un utilisateur lors de tâches comme l'apprentissage d'un logiciel ou la recherche d'informations dans des documents multimédia (web). Cela en vue de construire un modèle de l'utilisateur à partir d'informations verbales [FT95] et non-verbales recueilli lors de ces expériences.

Références

[ABA92] Mourad Abed, Jean-Marc Bernard, and Jean-Claude Angué. méthodes et moyens d'acquisition et de traitement des mouvements oculaires : application à la conception et à l'évaluation des interfaces homme-machine. In *Actes d'Interface des Mondes Réels et Virtuels*, pages 667–682, Montpellier, France, 1992.

[BFNT95] 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, Juin 1995.

[BKG95] 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.

[BP94] Shumeet Baluja and Dean Pomerleau. Non-Intrusive Gaze Traking Using Artificial Neural Networks. Technical report, ISL, Carnegie Mellon University, Pittsburgh, PA 15213, 1994.

[Bra96] Annelies Braffort. A gesture recognition architecture for sign language. In *Proc. of ASSETS'96, ACM*, Vancouver, Canada, Novembre 1996.

[Cad93] Claude Cadoz. Le geste canal de communication homme/machine. In *Cours de l'école d'été ARC/PRC CHM*, pages 35–67. 4–16 Juillet 1993.

[Cha95] 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, Octobre 1995.

[CM94] Colette Charbonnier and Dominique Massé. Écriture par commande visuelle. In *Actes d'Interface des Mondes Réels et Virtuels*, pages 185–191, Montpellier, France, 1994.

[FA93] Claudie Faure and Madeleine Arnold. L'interaction homme-machine du point de vue des principes d'économie. In *Actes d'IHM'93*, pages 3–8, France, 1993.

[FT95] Alain Finkel and Isabelle Tellier. From natural language to cognitive style. In *Proc. of 4th Intl. Colloque on Cognitive Sciences, ICCS*, San-Sebastian, Esp, 3–6 Mai 1995.

[GD96] 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, Juin 1996.

[Jac95] Robert J.K. Jacob. Eye Tracking in Advanced Interface Design. In W. Barfield and T.A. Furness, editors, *Virtual Environments and Advanced Interface Design*, pages 258–288. Oxford University Press, New York, USA, third edition, 1995.

[Mar95] Jean-Claude Martin. Types and goals of cooperation between media as quality criteria for multimedia interfaces. In *Proc. of the First International Workshop on Evaluation Methods and Quality Criteria for Multimedia Applications, ACM Multimedia 95*, San Francisco, USA, Novembre 1995.

[SYW96] 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.