Coopération de Techniques Sensorielles pour une Interaction Écologique

François Bérard, Joëlle Coutaz
CLIPS-IMAG
B.P. 53
38041 Grenoble Cedex 9, France
{francois.berard, joelle.coutaz}@imag.fr

RÉSUMÉ

La vision par ordinateur, de même que les techniques audio, offrent la possibilité de capter le comportement de l'utilisateur dans son milieu naturel sans adjonction d'artifices contraignants comme les cordons de connexion du gant numérique. Ces nouvelles capacités multi-sensorielles des traitements informatiques doivent, à notre sens, contribuer à une interaction écologique davantage respectueuse des conditions de travail de l'utilisateur. Dans cet article, nous analysons l'apport de la vision par ordinateur et de la capture audio pour deux domaines innovants de l'Interaction Homme-Machine : la réalité augmentée et la communication homme-homme médiatisée. Nous montrons que les limites des prototypes actuels sont dues à des techniques trop simplistes. Ces limitations nous amènent à présenter nos résultats comparatifs entre quatre techniques distinctes de suivi de l'utilisateur. Forts de cette expérience, nous proposons un processus coopératif de techniques de suivi dont la complémentarité fonctionnelle devrait offrir un suivi plus fiable, précis et autonome en vue d'une interaction réellement écologique.

MOTS CLÉS : Vision, Audition, Interaction Écologique, Communication Médiatisée.

INTRODUCTION ET MOTIVATION

En Interaction Homme-Machine, le développement de nouvelles techniques d'interaction constitue un thème de recherches actives. L'invention de la souris par Engelbart dans les années 60, symbole historique de ce courant d'étude, a suscité l'émergence de nouveaux paradigmes d'interaction, et notamment celui de la manipulation directe. Mais il faut admettre que les dispositifs d'interaction en général, tels l'écran tactile, le gant numérique et les systèmes à retour d'effort, ne sont pas toujours adaptés comme moyens d'action sur les machines.

Notre équipe travaille depuis deux ans sur le développement de nouvelles techniques d'interaction fondées sur les capacités sensorielles de la machine. L'objectif est de doter l'ordinateur, mais aussi l'environnement de travail, de dispositifs capables de capter le comportement des utilisateurs dans leur milieu naturel sans adjonction d'artifices contraignants. Nous visons tout mécanisme d'interaction qui élimine le "fil à la patte" des dispositifs actuels. Nous ne cherchons pas à éliminer l'acquis mais nous visons à le compléter. Dans l'état actuel de notre recherche, nous avons retenu l'audition et la vision par ordinateur comme pistes de réflexion mais notre intérêt va également vers d'autres types de capteurs comme les infrarouges ou les détecteurs de présence.

Notre recherche est motivée par le concept d'"interaction écologique", c'est-à-dire une interaction qui soit "naturelle" (au sens de Norman), intégrée et transparente (au sens de Buxton) pour une classe de tâches données. Sur ce point, la caméra vidéo et le microphone présentent des avantages déterminants sur les nouveaux dispositifs d'entrée comme les gants sensitifs ou à retour d'effort. L'un et l'autre peuvent s'insérer dans l'environnement sans imposer à l'utilisateur de harnachement spécifique. Ils savent se faire oublier tout en procurant une information riche et exploitable en interaction homme-machine.

Dans cet article, nous montrons l'intérêt du son et de la vision par ordinateur en tant que techniques de suivi pour la réalité augmentée et la communication homme-homme médiatisée. Nous observons que la qualité du suivi constitue un elément déterminant de la validité écologique de la vision par ordinateur. Nous présentons ensuite une étude comparative entre plusieurs techniques de suivi par vision que nous avons mises en œuvre et nous montrons comment l'information sonore peut également renseigner le système sur la position du locuteur. Les propriétés de ces diverses techniques sensorielles (audio et vidéo) nous amènent à formuler une nouvelle proposition en cours de mise en œuvre dans notre prototype de mediaspace CoMedi.

Vision pour la réalité augmentée

La réalité augmentée trouve une illustration claire de ses principes avec le concept Bureau Digital [6]. Le Bureau Digital marque aussi l'une des premières utilisations de la vision par ordinateur en tant que technique d'interaction homme-machine. Comme le montre la figure 1, le Bureau Digital est un bureau physique "augmenté" de services informatiques : les informations numériques sont projetées sur la table de travail tandis que les gestes de l'utilisateur sont captés par une caméra vidéo. L'idée centrale du Bureau Digital est de briser la frontière entre les fonctions et objets physiques et les fonctions et objets virtuels (numériques). Les fonctions informatiques sont intégrées à l'environnement de travail naturel de l'utilisateur créant ainsi une "réalité augmentée". Clavier, écran et

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France

souris disparaissent au profit des techniques usuelles de manipulation : le doigt, le stylo, la gomme, la table, le papier. Cet exemple met en évidence un requis de qualité écologique des dispositifs de manipulation. C'est tout naturellement la vision par ordinateur que Wellner a utilisée.

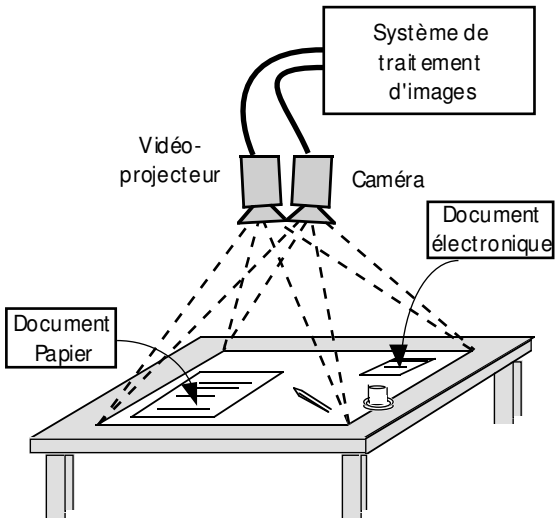

Figure 1 : L'installation du Bureau Digital.

Le prototype de Bureau Digital met en évidence plusieurs problèmes majeurs de vision par ordinateur [3]. Certaines fonctions comme l'identification de l'objet manipulé sont simulées. Seul le suivi du doigt est réalisé mais selon une technique rudimentaire : comme le montre la figure 2, deux images consécutives du flot vidéo sont soustraites l'une de l'autre faisant apparaître les zones de l'image qui ont bougé. On suppose ensuite que le pointeur (un doigt, un stylo) est à l'extrémité haute de la "zone de mouvement" mise en évidence par la différence d'images.

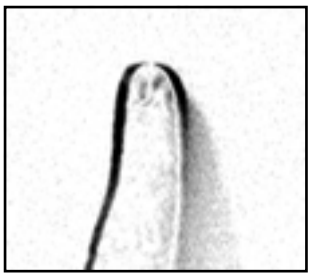

Figure 2 : Différence entre deux images successives du déplacement latéral d'un doigt. Noter la difficulté d'en déduire une position de pointeur dans le sens horizontal.

La différence d'images est une technique simple et peu coûteuse mais elle manque de précision : 1) le calcul de la position du pointeur s'appuie sur deux images successives donc sur deux positions différentes du pointeur. Le pointeur doit effectuer un déplacement non négligeable pour être détecté. La technique est donc incapable de suivre les micro-déplacements d'une désignation de précision. 2) le pointeur est supposé correspondre à l'extrémité haute de la zone de mouvement. Cette hypothèse s'avère inexacte dans certains cas : par exemple la pointe d'un stylo est plus basse dans l'image de différence que les articulations des phalanges de la main qui le tient.

Le prototype du Bureau Digital montre l'apport de la vision par ordinateur à l'interaction homme-machine. Cependant, la simplicité de la technique ne permet pas de satisfaire au critère d'interaction écologique : l'utilisateur est soumis à des contraintes imposées par le système (déplacements exagérés, maintien du pointeur "vers le haut"). Nous allons voir que cet obstacle se manifeste de façon similaire dans l'application de la vision par ordinateur à une autre domaine de l'IHM : la Communication Homme-Homme Médiatisée (CHHM).

Vision pour la communication homme-homme médiatisée

La CHHM désigne les fonctions de communication dans les applications de travail coopératif (collecticiels). Depuis le courrier électronique, les applications de CHHM ont évolué vers la visioconférence et plus récemment sous forme de mediaspace. Le mediaspace est apparu par contraste à la visioconférence qui suppose des réunions planifiées et protocolaires. Dans l'accomplissement de tâches formalisées, le média vidéo n'est pas nécessairement pertinent [5]. Inversement, il semble que le support vidéo joue un rôle déterminant sur la conscience de groupe en communication informelle et opportuniste comme dans les mediaspaces dont les services, accessibles en permanence, tentent de reproduire une "coprésence" physique et renforcent le sentiment d'appartenance à un groupe.

Cependant, le caractère permanent du mediaspace ne suffit pas aux objectifs de coprésence et d'informalité. Des progrès sont nécessaires et notamment au niveau fin de l'interaction entre les utilisateurs via les supports de communication. Par exemple, le point de vue dont l'utilisateur de mediaspace actuel dispose est limité au champ de vision de la caméra distante sur laquelle il n'a aucun contrôle (sous réserve, évidemment, que les droits d'accès soient respectés). Pour palier cette limitation, Gaver propose le concept de fenêtre virtuelle [4]. Le principe est de simuler l'effet d'un déplacement latéral face à une fenêtre du monde réel : le point de vue sur la scène distante change. Sur le plan technique, il aurait été aisé de piloter la caméra distante par l'intermédiaire d'un joystick ou d'une boîte de contrôle à l'écran. Ce choix n'est pas acceptable dans le contexte des mediaspaces car l'introduction de manipulations artificielles va à l'encontre d'une communication naturelle entre les personnes. De fait, Gaver s'est tourné vers la qualité écologique de la vision par ordinateur : la réalisation du concept de fenêtre virtuelle s'appuie sur le suivi de visage par vision par ordinateur. Lorsque l'utilisateur se déplace latéralement face au moniteur, son mouvement est détecté et permet de guider la caméra distante dans la direction opposée.

Comme pour le Bureau Digital, la réalisation de la fenêtre virtuelle repose sur une différence d'images.

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France

Toutefois, dans le cas de la fenêtre virtuelle, les images du flux vidéo sont soustraites à une image de référence plutôt qu'à l'image précédente. La technique nécessite une phase d'initialisation pendant laquelle l'image de référence est capturée lorsque l'utilisateur est hors du champ de la caméra. Pendant le suivi, la différence entre les images du flux vidéo et l'image de référence fait apparaître l'utilisateur (cf. figure 3).

Figure 3 : Image de référence, image analysée et image de différence seuillée.

Ce procédé a l'avantage de fournir la position de l'utilisateur même s'il ne bouge pas. Par contre, Gaver relève plusieurs défauts dûs au manque de robustesse et de précision du suivi. Par exemple, l'image de référence doit est régulièrement mise à jour pour compenser les variations lumineuses, imposant à l'utilisateur de sortir du champ de la caméra. Le manque de robustesse implique des comportements parasites sur l'utilisateur : une fois qu'un bon point de vue est affiché, les personnes qui ont testé le système renoncent à bouger de peur de ne pas pouvoir le retrouver. Enfin, de par son principe, la technique impose un point de vue fixe de la caméra locale (sans quoi l'image de référence n'aurait plus de sens). Or il est souhaitable que l'utilisateur local puisse se déplacer librement et donc que la technique de suivi fonctionne également pour une caméra active.

En résumé, que ce soit pour le Bureau Digital ou la fenêtre virtuelle, les études menées jusqu'ici montrent qu'il y a avantage pour l'utilisateur à disposer de capteurs écologiques. Sur le plan technique, l'écologie de ces capteurs implique des algorithmes de traitement précis et robustes mais complexes. Nous pourrions concevoir de nouveaux algorithmes. Nous tentons une autre approche fondée sur l'acquis. Il s'agit d'étudier, voire affiner, les techniques actuelles, d'en identifier les caractéristiques et de les faire coopérer en jouant sur leurs compétences fonctionnelles respectives. C'est ce que nous présentons dans les sections qui suivent.

ÉTUDE COMPARATIVE DE QUATRE TECHNIQUES SENSORIELLES POUR LE SUIVI DU VISAGE

Nous avons étudié quatre techniques de suivi distinctes : la corrélation, l'histogramme de couleur, la détection du clignement des yeux et la localisation de la source sonore. Nous parlerons de *suivi* lorsque la position précédente de la cible est utilisée pour la recherche de la nouvelle position, alors que le terme *localisation* sera employé lorsque la cible est recherchée de manière globale dans chaque image. Toutefois ces techniques ont toutes la même finalité : connaître la position de la cible à tout instant.

Le prototype sur lequel nous travaillons est un mediaspace augmenté de capacités sensorielles. C'est pourquoi les techniques utilisées sont destinées à suivre les visages mais la corrélation et l'histogramme de couleur sont applicables à d'autres cibles.

Suivi par Corrélation

Le suivi par corrélation est étudié depuis presque 10 ans dans le domaine de la vision par ordinateur [1]. Cette technique a suscité de nombreuses études théoriques et expérimentales [5] dont les résultats nous procurent une bonne maîtrise de ses paramètres. Le principe, illustré par la figure 4, est le suivant : étant donné un motif (petite zone de l'ordre de 16x16 pixels) dont on connaît la position dans l'image i, trouver sa nouvelle position dans l'image i+1. Pour cela, on définit une zone de recherche dans l'image i+1 qui correspond à un voisinage (de l'ordre de 32x32 pixels) de la position du motif dans l'image i. Le motif est alors comparé à toutes les zones de même taille qu'il est possible de définir dans la zone de recherche. La position de la zone qui ressemble le plus au motif est choisie comme nouvelle position du motif dans l'image i+1.

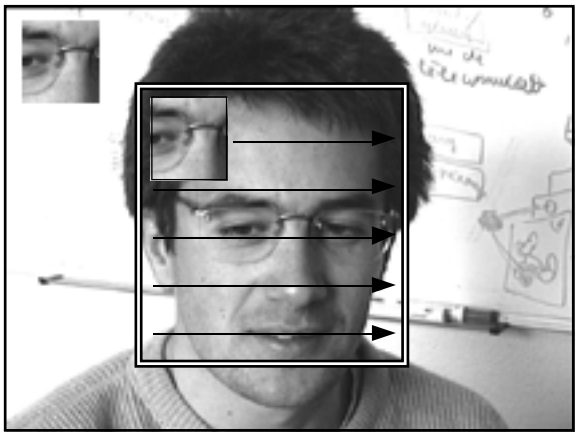

Figure 4 : Suivi par corrélation.
Le motif, représenté en haut à gauche, est recherché à l'intérieur d'un voisinage (carré blanc) de sa position dans l'image précédente.

Le suivi par corrélation est stable et précis (le motif est localisé au pixel près). Il présente cependant trois points faibles : 1) la localisation du motif dans la zone de recherche s'appuie sur les valeurs des pixels. Ces valeurs varient fortement en fonction de l'illumination de la cible. La corrélation est de fait inapte à suivre une cible qui passe d'une zone claire à une zone d'ombre. 2) L'algorithme de recherche du motif est très coûteux. Le remède consiste à limiter la taille de la zone de recherche mais alors le suivi échoue lorsque la cible se déplace trop rapidement. Il faut donc jouer sur les compromis taille motif, taille recherche (par exemple sur Macintosh Quadra 700, un motif de 8x8 pixels et une zone de recherche de 20x20 pixels autorisent un suivi à 15 Hz). 3) La technique nécessite une phase d'initialisation pour choisir un motif sur une zone de l'image qui représente la cible du suivi.

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France

Localisation par Histogramme de couleurs

L'expression "histogramme de couleurs" est un abus de langage : la couleur d'un pixel définit son aspect perceptible à l'écran (souvent représenté par trois composantes rouge, vert et bleu). Dans le cas de l'histogramme, l'algorithme utilise la teinte des pixels, c'est-à-dire la couleur du pixel normalisée par son intensité lumineuse. Deux pixels de même teinte ont des couleurs différentes lorsqu'ils représentent une zone claire et une zone sombre de l'image. Ainsi, un algorithme utilisant la teinte des pixels est plus robuste aux variations d'intensité lumineuse que s'il utilisait leur couleur.

Le principe du suivi par histogramme de couleurs consiste à identifier les pixels de l'image qui s'approchent le plus d'une teinte donnée. Dans le cas du suivi du visage, la teinte de la peau sert de référence. Dans une phase de calibrage, l'algorithme analyse un échantillon de pixels représentant une partie du visage. Il en ressort les valeurs définissant la teinte de la peau. Dans la phase de localisation, chaque pixel de l'image est remplacé par la valeur de ressemblance de sa teinte à la teinte de la peau. Comme le montre la Figure 5, l'image obtenue, appelée image de probabilité, fait apparaître le visage dans ses zones claires.

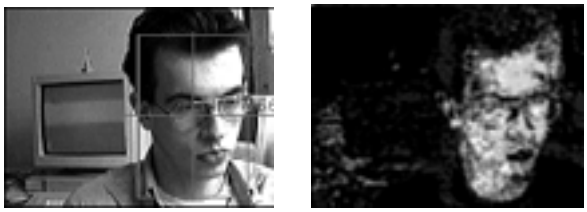

Figure 5 : Image de départ et image de probabilité. Sur cette dernière les pixels clairs sont ceux dont la teinte est proche de celle de la peau.

Parmi les atouts de cette technique, on relève l'efficacité et la généralité. La recherche peut se faire sur l'ensemble de l'image à une fréquence raisonnable (6 Hz à une résolution de 160x120 sur un Power Macintosh) ; cette technique autorise la localisation de plusieurs visages (voire tout objet) dans le champ de la caméra sans surcoût : à chaque visage est associé une tâche claire dans l'image de probabilité.

Inversement, nous avons constaté que la détermination de la teinte des pixels imposait un bon éclairage de la scène, ce qui tempère l'indépendance de la technique vis-à-vis des conditions lumineuses.

Localisation par clignement des yeux

Le clignement des yeux est une action naturelle inconsciente dont on peut tirer profit pour localiser le visage [2]. La technique revient à calculer continuellement la différence entre deux images du flux vidéo séparées de 0,1 secondes (demie-durée d'un clignement). Si les déplacements du sujet qui fait face à la caméra sont faibles, l'image de différence obtenue est presque entièrement blanche : deux images séparées d'un si court instant sont presque identiques et leur différence est nulle. Cependant, au moment où un clignement des yeux intervient, il se forme deux tâches sombres au niveau de la position des yeux. Un exemple en est donné sur la figure 6. Après plusieurs traitements sur l'image de différences (détaillés dans [2]), la position des yeux est localisée, ce qui permet d'en déduire la position du visage.

Cette technique présente trois atouts importants : elle ne nécessite pas d'initialisation ; elle est peu coûteuse du fait de la simplicité de ses algorithmes. En conséquence, elle autorise une recherche dans tout le champ de l'image à une fréquence de fonctionnement acceptable (nous l'avons expérimentée à 12 Hz avec une résolution de 160x120 sur un Macintosh Quadra 700). En outre, cette technique s'appuie sur l'aspect dynamique du flot vidéo, ce qui la rend bien plus robuste aux variations lumineuses que les techniques qui analysent l'image de manière statique.

Figure 6 : Deux images capturées pendant un clignement des yeux, et l'image de différence résultante.

La contrepartie vient de sa disponibilité : la localisation n'est possible que lors des clignements des yeux dont la fréquence est très variable. En conséquence, elle n'autorise pas un suivi continu.

Localisation de la source sonore

Considérons une personne parlant face à un écran équipé d'un microphone à chacun de ses quatre angles, comme schématisé sur la figure 7.

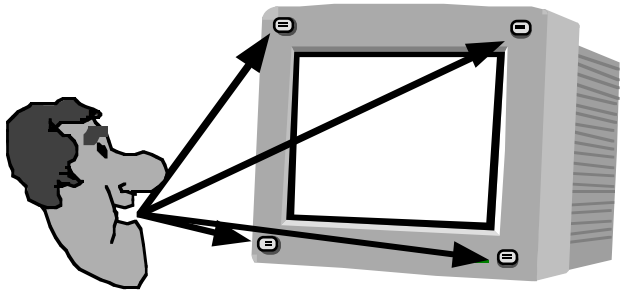

Figure 7 : Localisation de la source sonore. Les différences de distance entre les micros et la source permettent de la localiser.

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France

	Corrélation	Clignement des yeux	Histogramme de couleur	Localisation de la source sonore
Portée	+ (région de l'image)	+++	+++	+++
Précision	+++	+++	++	+
Stabilité	+++	++	++	+
Tolérance aux variations lumineuses	+	+++	++	+++ (non affecté)
Disponibilité	+++	+	+++	++
Initialisation nécessaire	Oui	Non	Oui	Non
Atout spécifique			multicible sans surcoût	identifie le locuteur <=> centre d'intérêt

Figure 8 : Tableau récapitulatif des caractéristiques des quatre techniques de suivi du visage étudiées.

Les signaux sonores captés par les micros sont presque identiques, à cela près qu'ils sont décalés dans le temps : par exemple le signal reçu par un micro m, dont la distance à l'utilisateur est supérieure à celle du micro n, présente un retard par rapport au signal capté par le micro n. En comparant entre eux les quatre signaux, il est possible de mesurer précisément ces écarts. On en déduit les différences de distance entre l'utilisateur et les quatre micros, puis, par résolution d'un système, la position dans l'espace de la bouche de l'utilisateur.

Avec cette technique, nous sommes confrontés à un problème similaire à celui du clignement des yeux : l'information de position n'est disponible que lorsque l'utilisateur parle. En particulier, il serait vain de vouloir localiser tous les participants d'une même salle pendant une visioconférence car il n'y a, en général, qu'une seule personne qui parle. D'un autre côté, nous bénéficions ici d'une information essentielle : là où les techniques de vision ne perçoivent qu'un ensemble de visages indifférenciés, la localisation de la source sonore identifie la personne qui parle et qui incarne, en général, le centre d'intérêt.

Bilan

Le tableau de la figure 8 présente un bilan synthétique des caractéristiques des quatre techniques étudiées. A l'évidence, on n'y relève pas de solution "miracle" mais les atouts des unes peuvent servir à combler les lacunes des autres. Au vu du tableau, la coopération de techniques de suivi offre une voie de recherche prometteuse. C'est la stratégie que nous allons adopter pour notre prototype de mediaspace "CoMedi".

COMEDI : UN MEDIASPACE SENSORIEL

Le projet CoMedi (Communication et Médiaspace) offre des services accessibles en permanence de télé coup-d'œil, de télévisite et de visioconférence. Il doit servir de terrain d'expérience à l'étude de solutions techniques mais aussi à l'analyse du comportement des utilisateurs (protection de l'espace privé, perception de nos capteurs écologiques). CoMedi va être augmenté d'une coopération des techniques sensorielles présentées ci-dessus pour offrir :

- un suivi local du locuteur : la caméra locale, montée sur tourelle, suivra l'utilisateur pour qu'il n'ait pas à se soucier de rester dans le champ de la caméra ;
- le mécanisme de la fenêtre virtuelle pour pratiquer des tours d'horizon à distance.

La coopération entre les techniques de suivi comprend deux facettes complémentaires : fusion et contrôle. Concernant la fusion, chaque technique impliquée dans le suivi fournit son estimation de la position du visage. Tout résultat est fusionné dynamiquement à la connaissance courante de la position du visage. Pour cela, nous avons défini une "interface" de programmation commune à tous les modules de suivi. Au retour d'un appel, chaque module renvoie trois paramètres : la position estimée du visage (coordonnées x, y et éventuellement z), la covariance de cette position (borne estimée de l'erreur sur la position), et le facteur de confiance (taux de validité estimé des deux premiers paramètres). La fusion de ces données, application directe de résultats mathématiques, devrait fournir une estimation de la position du visage plus précise et plus stable que celle des modules pris isolément.

La seconde facette de la coopération concerne le contrôle mutuel des techniques. Un super-contrôleur est en cours de mise en œuvre. Il s'agit essentiellement d'un graphe d'états dont chaque nœud représente l'appel à un module de suivi ou bien au module de fusion ou encore au module de pilotage de la caméra.

Ce projet est en cours de réalisation sur stations de travail Silicon Graphics Indy, équipées de caméra pilotables Canon VC-C1. Les traitements des images sont réalisés en C++, alors que le contrôleur est développé en tcl/tk. Ce langage interprété devrait faciliter la mise au point expérimentale du graphe du contrôleur.

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France

CONCLUSION

L'introduction en Interaction Homme-Machine de techniques sensorielles comme l'audition et la vision par ordinateur ouvrent la voie à de nouvelles formes d'interaction dont le caractère écologique était jusqu'ici irréalisable. Leur conception pose cependant de vraies difficultés : les premières tentatives d'interprétation du flux de données capté ont conforté notre motivation pour le développement de ces techniques mais ont également montré leurs limites.

Dans cet article, nous décrivons trois techniques de vision et une technique d'audition en mettant en évidence leurs forces et faiblesses. Le bilan de cette étude nous amène à proposer un processus de coopération jouant à la fois sur la redondance d'information et la complémentarité fonctionnelle. Le premier aspect devrait apporter précision, stabilité et robustesse à l'interaction produite, alors que le second lui donnera son autonomie.

Si les technique mises en jeux existent déjà isolément, il nous reste à réaliser le processus de coopération permettant de valider notre proposition. Dans la perspective d'un succès technique, c'est-à-dire l'exécution d'un processus de suivi fiable et autonome, il nous restera à valider notre intuition sur l'apport de l'interaction écologique en IHM. C'est dans ce but que nous comptons ensuite mener une étude expérimentale d'utilisabilité de notre mediaspace sensoriel.

REMERCIEMENTS

Ce travail a été partiellement financé par le projet ESPRIT AMODEUS et reçoit le soutien de France Télécom CNET. Nous tenons à remercier Yves-Henri Berne et Bruno Hocq pour le développement de la maquette de suivi qui a servi de démonstrateur de faisabilité pour CoMedi.

BIBLIOGRAPHIE

1. Anadan P. *"Measuring Visual Motion From Image Sequence"*. PhD dissertation and COINS Technical Report 87-21, University of Massachussetts, Amherst, 1987.

2. Bérard F. *"Vision et IHM : Application à l'analyse des visages"*. Mémoire de Magistère, septembre 1993, Université Joseph Fourier, Grenoble.

3. Bérard F. *"Vision par ordinateur pour la réalité augmentée"*. Mémoire de DEA, juin 1994, Université Joseph Fourier, Grenoble.

4. Gaver W., Smets G., et Overbeeke K. *"A Virtual Window On Media Space"*. Actes de la conférence CHI'95, ACM press, pp. 257-264.

5. Martin J., Crowley J.L. *"Experimental Comparison of Correlation Techniques"*, IAS-4, International Conference on Intelligent Autonomous Systems, Karlsruhe, Mars 1995.

5. Sellen A. *"Remote Conversations : The Effect of Mediating Talk With Technology"*. Human-Computer Interaction, 1995, Volume 10, pp. 401-444.

6. Wellner P. *"Interacting with paper on the DigitalDesk"*. Communications of the ACM, July 1993, Vol.36 No.7, pp 87-97.

Publié dans les actes des huitièmes journées sur l'Interaction Homme-Machine,
Septembre 1996, Grenoble, France