Big Brother is watching but helping you : analyse et interprétation de mouvements humains (expressions, gestes, postures)

Alice Caplier

Big Brother is watching but helping you : analyse et interprétation de mouvements humains (expressions, gestes, postures)

Mémoire HDR version 1.0

Alice PASCAL CAPLIER

Soutenue le 15 décembre 2005

Jury :

- **Pierre-Yves COULON, PR INPG, LIS (président)**
- **Michel DHOME, DR CNRS, LASMEA (rapporteur)**
- **Maurice MILGRAM, PR Paris 6ème, LISIF (rapporteur)**
- **Ferran MARQUES, PR Université Polytechnique de Catalogne, Espagne, (rapporteur)**
- **Laurence NIGAY, PR UJF Grenoble, CLIPS (examinatrice)**
- **Pascal PERRET, ingénieur France Télécom R&D Meylan**

Table des matières

Liste des figures

Partie I : CV et résumé du dossier

1 Curriculum vitae

1.1 Etat civil

Nom patronymique :	PASCAL
Nom marital :	CAPLIER
Prénoms :	Alice Anne Mireille
Date et lieu de naissance :	25 juin 1968 à Paris
Nationalité :	française
Situation de famille :	mariée, 1 fille née en 1995, 1 fils né en 2000
Adresse professionnelle :	Laboratoire des Images et des Signaux (LIS), INP Grenoble- 46 avenue Félix Viallet, 38031 Grenoble Cedex.
Numéro de téléphone :	04-76-57-43-63
Fonction :	Maître de Conférences
Etablissement actuel :	Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG) / Institut National Polytechnique de Grenoble (INPG)

1.2 Diplômes universitaires

- 1991 : diplômée de Ecole Nationale Supérieure des Ingénieurs Electriciens de Grenoble.
- 1992 : DEA **S**ignal **I**mage **P**arole de l'INPG.
- 1995 : Thèse de doctorat de l'INPG soutenue le 20/12/95.
  Sujet : Modèles markoviens de détection de mouvement dans les séquences d'images : approche spatio-temporelle et mises en oeuvre temps réel.

1.3 Parcours professionnel

- 1992/1995 : Thèse de doctorat au LTIRF (Laboratoire de traitement d'images et de Reconnaissance de formes).
- 1995/1997 : ATER : enseignement à l'ENSERG et recherche au LTIRF
- depuis 1997 : maître de Conférences : enseignement à l'ENSERG et recherche au Laboratoire des Images et des Signaux

Rque : le LIS a été créé en janvier 97 par fusion du LTIRF et du Cephag (Centre d'étude des phénomènes aléatoires de Grenoble).

1.4 Rayonnement national et international

- J'ai été membre du Conseil de Laboratoire du LIS pendant 6 ans ;
- Je suis membre du Conseil d'administration de l'ENSERG ;
- Je suis membre suppléant de la commission paritaire de l'ENSERG ;
- Je suis membre de la CSE 61ème section de l'INPG (second mandat en cours);
- J'ai été membre suppléant de la CSE 61ème section de l'université de Lyon pendant 4 ans;
- Je suis membre du Conseil Scientifique de l'INPG (premier mandat en cours).

- J'ai participé à l'organisation de l'atelier *Acquisition du geste humain par vision artificielle et applications* soutenu par l'AS 70 du CNRS.
- J'ai été relecteur pour les conférences *ICPR 2002, EUSIPCO 2005* ainsi que pour les revues *IEEE Transactions on Circuits and Systems for Video technology*, *Journal of Electronic Imaging, Eurasip Journal on Signal Processing.*

2 Activités de recherche

2.1 Résumé

Mes activités de recherche portent sur l'analyse et l'interprétation des mouvements humains à partir de données visuelles avec comme application principale l'amélioration du processus de communication entre l'homme et la machine. L'idée sous-jacente est de tendre vers une communication homme machine non pas par l'intermédiaire des traditionnels écran/clavier/souris mais vers un processus plus « humain » de communication. Ceci suppose que la machine est capable de reconnaître et d'interpréter tous les signes de communication humaine à savoir le langage verbal mais aussi tous les signes de communication non verbale. Nous nous plaçons donc en amont de la problématique des interactions homme-machine proprement dites. Ces travaux se focalisent sur l'interprétation des gestes humains de communication non verbale et en particulier sur l'interprétation des expressions faciales, des mouvements de la tête rigides (hochements…) ou non rigides (clignements, bâillement…), de certains gestes de mains (langage parlé complété destiné aux malentendants) ainsi que de certaines postures (assis, debout…). Pour tous les gestes considérés, la méthodologie de reconnaissance ou interprétation se déroule en deux phases :

- Dans un premier temps, nous développons des méthodes performantes d'extraction d'informations de bas niveau (qui ne possèdent pas de signification sémantique). Plusieurs types d'informations bas niveau sont extraites : des contours, des régions et des informations de mouvement. Pour l'extraction de contours de traits du visage tels que les yeux, la bouche et les sourcils, nous avons développé des algorithmes utilisant des modèles paramétriques de contours adaptés aux formes recherchées. Les paramètres des modèles de contours de type courbes de Bézier ou courbes cubiques sont estimés par des méthodes robustes de maximisation de flux de gradient de luminance et/ou de chrominance. Par ailleurs, afin d'extraire et de suivre les régions mobiles présentes dans une scène, nous avons développé et amélioré une méthode de détection de mouvement basée sur une modélisation markovienne associée à des observations de différence d'images et de différence avec une image de référence. Le suivi des régions extraites se fait par utilisation d'un filtre de Kalman à données éventuellement incomplètes dans le cas où la scène présente des occultations des objets mobiles. De plus, afin d'estimer automatiquement la nature et la direction des mouvements associés à la tête d'une personne, nous avons développé une méthode fréquentielle efficace qui s'appuie sur la modélisation des traitements qui se produisent au niveau de la rétine humaine. Enfin, une méthode de segmentation de la main s'appuyant sur des informations de chrominance et sur une carte de transformée de distance permet de tendre vers un modèle de main avec une probabilité de présence associée à chaque doigt. En effet, la classification des gestes du langage parlé complété est assez simple dès lors que l'on a réussi à identifier le nombre de doigts dépliés sur la main codeuse.
- Dans un second temps, une phase de fusion de données est envisagée afin d'aboutir à une interprétation de haut niveau ayant une signification sémantique (exemples : expression présente sur un visage, hochement d'approbation ou de négation, état de

somnolence, personne couchée ou assise…). Bien que des méthodes de type HMM ou réseau bayesien aient été considérées à des fins de comparaison de performances de classification, la principale méthode de fusion de données qui a été utilisée est la théorie de l'évidence. En effet, vis à vis des problèmes de classification de postures ou de classification d'expressions faciales, nous estimons que cette approche présente l'avantage de pouvoir considérer des mélanges d'expressions ou des mélanges de postures et d'autre part, elle permet la prise en compte explicite de données imprécises ce qui est important ici puisque toutes les données résultent de méthodes de traitement d'images. Nous nous penchons en particulier sur le problème de voir comment étendre cette théorie à un cadre de classification dynamique (prise en compte d'informations spatiales et temporelles).

Toutes ces activités de recherche ont été menées en collaboration avec 5 thésards (voir 2.2.1), 2 étudiants de DRT (voir 2.2.2) et 11 étudiants de DEA (voir 2.2.3). Elles ont été développées dans le cadre de plusieurs projets de recherche nationaux (projet RNRT TempoValse, projet BQR Telma, projet PEGASUS Deixis, projet Présence du cluster régional « Informatique, signal et logiciels embarqués ») et européens (projet Art-live, réseau d'excellence Similar). Elles se sont accompagnées de 10 articles dans des revues à comité de lecture, 35 articles dans des conférences internationales et 9 articles dans des conférences nationales. La liste complète des références associées à ces publications est donnée à la fin du mémoire.

2.2 Encadrement de recherche

Le Directeur de Thèse

Pour commencer une thèse, il faut avoir un patron. Un patron, c'est un monsieur très, très fort qui me pose un problème et qui va m'aider à le résoudre.
Des fois aussi, ça se passe mal, parce que je me trompe. Et quand je me trompe, avec mon patron, ça ne rigole pas, mais alors pas du tout. 'Regardez-moi dans les yeux, Nicolas', il me dit, pas content du tout. 'Vous appelez ça du travail, peut-être ?' qu'il me demande. Eh ben, là, ça a l'air d'une question, mais il ne faut surtout pas répondre, parce que sinon, il se fâche tout rouge !
(extrait : le Petit Nicolas en thèse, http://maesa.95mb.com/nicolas.htm)

2.2.1 Encadrement de thèse

1. Nicolas EVENO - Segmentation des lèvres par un modèle déformable analytique -, allocataire AMN, co-encadrement à 50% avec P.Y. Coulon (50%), **soutenue** le 14 novembre 2003, *thèse de l'INPG, spécialité Signal, Image, Parole*, laboratoire LIS, Grenoble.
2. Vincent GIRONDEL -Analyse et interprétation du mouvement humain pour des applications de réalité mixte- allocataire BDI, co-encadrement à 45% avec L. Bonnaud (45%) et J.M. Chassery (10%), **soutenance prévue** à l'automne 2005.
3. Zakia HAMMAL - Extraction dynamique des traits caractéristiques du visage en vue de la reconnaissance d'émotions- bourse Egide, co-encadrement à 90% avec Jeanny Hérault (10%) **soutenance prévue** à l'automne 2005.

4. Alexandre BENOIT - Extraction d'informations sur la posture en vue de la reconnaissance d'expressions faciales- , allocataire du ministère, encadrement à 100% (agrément obtenu à l'automne 2003), **soutenance prévue** à l'automne 2006.
5. Thomas BURGER – Segmentation temps réel de vidéos de gestes du langage parlé complété- Thèse Cifre France Télécom R&D, encadrement à 100% (agrément obtenu à l'automne 2004), **soutenance prévue** à l'automne 2007.

2.2.2 Encadrement de DRT

J'ai obtenu un agrément pour l'encadrement de chacun de ces deux DRT :

1. Antoine ROBINET, DRT micro-électronique et traitement du signal, **soutenu** en septembre 2000 - Contrôle de trafic routier par un micro-système magnétométrique.- Travail réalisé au LETI, CEA Grenoble.
2. Sandrine PIRES DRT Electronique, Electrotechnique, Automatique, **soutenu** en octobre 2002, -Détection automatique de défauts dans des images radiographiques obtenues par tomosynthèse tangentielle.- Travail réalisé au LETI, CEA Grenoble.

2.2.3 Encadrement DEA

1. D.Lam : Détection de contours par modèles de formes actifs. Application à la détection de lèvres. DEA Signal Image Parole de l'INPG, 1998.
2. N.Eveno : Détection du contour des lèvres dans des images de visages parlants acquises avec différents cadrages de la caméra. Interpolation d'images. DEA Signal Image Parole de l'INPG, 2000.
3. N. Mottin : Localisation de visages dans des images acquises avec différents cadrages de caméra. Application à l'indexation. DEA Algorithmique, Robotique, Automatique, Vision, Image et Signal de Sophia-Antipolis, 2000.
4. A. Haddad : Etude d'un algorithme de contours actifs pour la segmentation des lèvres. DEA Signal Image Parole de l'INPG, 2001.
5. V. Girondel : Détection de peau, suivi de têtes et de mains pour des applications multi-média. DEA Signal Images Parole de l'INPG, 2002 (co-encadrement avec L. Bonnaud).
6. Y. Lauriou : Détection de personnages. Application à l'indexation d'images. DEA Signal Image Parole de l'INPG, 2002 (co-encadrement avec A. Guérin).
7. Z. Hammal : Détection et localisation des yeux et des sourcils dans une séquence vidéo. DEA Intelligence artificielle et algorithmique de l'université de Caen, 2002.
8. J. Romeuf : Suivi de marqueurs dans des séquences vidéo de mouvement humain. DEA Signal Image Parole de l'INPG, 2003 (co-encadrement avec L. Bonnaud).
9. N. Cacciagguera : Calcul de paramètres et apprentissage de modèles de corps humain. DEA SiCom de l'université de Nice-Sophia Antipolis, 2003 (co-encadrement avec L. Bonnaud).
10. C. Piacenza : Estimation de la direction pointée par la main. DEA Signal Image Parole de l'INPG, 2004 (co-encadrement avec L. Bonnaud).
11. Y. Lahaye : Estimation de la direction du regard. DEA Signal Image Parole de l'INPG, 2005, (co-encadrement avec L. Bonnaud).

3 Activités d'enseignement

Une bonne partie de mon temps est également consacrée à l'enseignement à des étudiants destinés à devenir ingénieur soit par la filière classique des écoles d'ingénieur de l'Institut National Polytechnique de Grenoble (Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble, Ecole Nationale Supérieure des Ingénieurs Electriciens de Grenoble, Département Télécoms, Ecole Nationale Supérieure de Physique de Grenoble), soit

par des filières différentes telles que la formation continue de l'INPG ou le Centre Universitaire d'Etude et de Formation pour Adultes. La multiplicité des lieux où j'enseigne me permet de rencontrer des auditeurs de formations antérieures différentes ce qui est très intéressant.
Je suis amenée à effectuer des enseignements divers dont certains ont un lien étroit avec mes activités de recherche. J'effectue ces enseignements sous forme de cours, TD ou TP.

3.1 Enseignements dispensés

3.1.1 Par le passé

- Cours et TD d'algorithmique à l'ENSERG (1ère année) : l'objectif est d'une part d'apprendre aux étudiants à structurer leur pensée lors de l'élaboration d'un programme informatique et d'autre part, de les initier au langage C.
- T.D. d'automatique à l'ENSERG (1ère et 2ème années) : ces T.D. d'automatique portent sur les thèmes des asservissements continus linéaires et non linéaires, de la représentation d'état des systèmes dynamiques linéaires et des systèmes asservis linéaires échantillonnés.
- T.P. d'électronique à l'ENSERG (1ère et 2ème année) : l'objet de ces T.P. intitulés de manière générique « TP d'électronique » est d'encadrer des manipulations permettant une illustration directe et complémentaire des cours d'électronique, d'automatique et de traitement du signal.
- Atelier d'électronique à l'ENSERG (1ère année) : il s'agit d'encadrer des séances au cours desquelles les étudiants élaborent et testent des montages électroniques analogiques et/ou numériques élémentaires tels que des montages de mesures de tensions alternatives, de mesures de fréquence ou de capacité ou des montages de générateurs de fonctions, de base de temps d'un oscilloscope

3.1.2 Actuellement

- Cours d'électronique numérique au Cycle Préparatoire Polytechnique de Grenoble (2ème année) : ce cours est destiné à donner aux étudiants les bases nécessaires à l'électronique numérique. Il est présente des notions relatives à l'algèbre de boole, la logique combinatoire et séquentielle, les circuits arithmétiques, la mise en œuvre et la technologie ainsi que sur les convertisseurs.
- Cours Mouvement-Vidéo-codage à l'ENSIEG (3ème année) et Master recherche SIPT : ce cours est constitué de deux parties. La première portant sur l'analyse du mouvement et vision par ordinateur, présente aux étudiants un certain nombre de connaissances sur les développements récents concernant l'analyse du mouvement dans les séquences d'images et les applications associées. La seconde partie effectuée par Laurent Bonnaud, enseignant à l'Université Pierre Mendès France de Grenoble et chercheur au LIS porte sur les normes de compression vidéo.
- Cours de compression audio-vidéo à l'ENSERG (3ème année) et Master recherche SIPT : ce cours présente les techniques de compression des images et des vidéos ainsi que les normes de compression vidéo les plus courantes (MPEG2, H264…). La partie relative à la compression des signaux audio est assurée par Laurent Girin, enseignant à l'ENSERG et chercheur à l'Institut de la Communication Parlée.
- Cours de traitement du signal à la formation continue de l'INPG : ce cours présente les bases du traitement du signal analogique et numérique sous un angle aussi applicatif que possible. En effet, ce cours s'adresse à des auditeurs de la formation continue dont l'objectif est de voir avant tout le lien entre les fondements du traitement du signal et

certaines applications concrètes qu'ils peuvent rencontrer dans le cadre de leur activité professionnelle. Un exemple concret concerne la compréhension du fonctionnement d'un analyseur de spectre analogique et d'un analyseur de spectre numérique.

- T.D. Traitement du signal au département Télécom de l'INPG (2ème année) : ces T.D. portent sur l'analyse des signaux continus, le filtrage, l'analyse des signaux discrets, la quantification et l'analyse des signaux aléatoires.
- T.P. de traitement du signal au département Télécom (2ème année) : les manipulations proposées dans le cadre de ces T.P. portent sur le thème général des modulations dans les télécommunications.

3.1.3 Pour l'avenir proche

Lors de l'année universitaire 2005-2006, je vais dispenser deux nouveaux enseignements :

- cours de traitement d'images à l'ENSPG (3ème année) : ce cours a pour objectif de proposer un panorama des techniques de base d'analyse des images statiques et des séquences vidéos (segmentation, extraction de contours, analyse de mouvement, compression d'images, amélioration d'images). Ces méthodes sont illustrées par la présentation de systèmes dédiés à des cas concrets (vidéo-surveillance, segmentation des contours d'un visage…)
- BE de traitement d'images à l'ENSPG (3ème année) : six manipulations de 3 heures sont prévues pour illustrer le cours et faciliter son assimilation.

3.2 Responsabilités pédagogiques

Je suis actuellement responsable des trois plate-formes de TP suivantes :

- responsabilité des TP d'électronique 2A à l'ENSERG : 16 manipulations différentes
- responsabilité des TP de traitement du signal au département Télécoms 2A : 5 manipulations différentes
- responsabilité de deux U.V. de TP d'électronique pour la formation DEST du CUEFA.

Dans le cadre de ces responsabilités de TP, il s'agit de :

- veiller à l'organisation (planning, rotations des étudiants, mise en place des examens) et au déroulement des séances tout au long de l'année,
- faire évoluer ou remplacer les manipulations de TP en fonction des évolutions technologiques (par exemple, proposer aux étudiants l'utilisation de matériel de mesure récents ainsi que de logiciels largement répandus dans le monde industriel),
- gérer les budgets investissement et fonctionnement associés.

L'Enserg, suite à une incitation de la part de l'INPG, s'investit dans le développement de nouvelles technologies d'enseignement. J'ai à ce titre été volontaire pour être membre du groupe de travail sur les TICE. Dans ce groupe, les objectifs dans un premier temps modestes consistent à créer des bases de données de sujets et de corrigés d'examen. Dans un second temps, nous travaillons sur le développement de QCM pour l'évaluation ou l'auto-évaluation des étudiants. D'un point de vue personnel, je suis tout à fait intéressée par les expériences faites en terme de nouvelles techniques d'enseignement. A ce titre, j'ai déjà effectué un cours de traitement du signal à distance (cours accessible sur le Web + séance de tutorat à distance) dans le cadre de la formation ELAN de l'INPG.

L'Enserg vient d'obtenir l'habilitation pour l'ouverture lors de l'année universitaire 2005-2006 d'une année spéciale « Traitement du signal » dont j'assure la responsabilité. Mon rôle consiste à aider et orienter les étudiants afin qu'ils se composent une année d'apprentissage cohérente et en accord avec leurs objectifs professionnels. En effet, le nombre d'étudiants

dans ce type de formation étant souvent restreint, il est tout à fait concevable de leur proposer un programme « à la carte ».
Enfin, une grande réforme de l'INPG (dont l'objectif est l'amélioration de la visibilité du point de vue de l'extérieur) est en marche pour la rentrée 2006 avec une restructuration des 10 écoles d'ingénieurs de l'institut. De ce fait, je suis amenée à participer à un certain nombre de réunions de réflexion autour de cette réorganisation.

Partie II : Synthèse des activités de recherche

4 Introduction

4.1 Interprétation du mouvement humain

Le thème de recherche qui est développé dans ce qui suit est relatif à l'analyse et à l'interprétation du mouvement humain sur la base d'informations vidéo (« looking at people domain » selon la terminologie anglaise [Gravilla99]). Plus précisément, il s'agit d'identifier et de reconnaître les actions du corps humain dans son ensemble (analyse de postures et de comportement) ou de certaines parties du corps humain (reconnaissance de gestes, analyse des expressions faciales, reconnaissance de mouvement de tête). Depuis une quinzaine d'années, ce thème de recherche est étudié de manière intensive tant au niveau national qu'au niveau international. En effet, certaines conférences centrées exclusivement sur ce domaine ont vu le jour. On peut citer entre autre *IEEE International conference on Automatic Face and Gestures Recognition* dont la septième édition aura lieu en 2006 ou encore le *workshop HAREM 2005* (*Human Activity Recognition and Modelling*). Au niveau national, le RTP 25 du département STIC du CNRS a soutenu en 2004 une action spécifique (AS 70) portant sur la *Perception, modélisation et interprétation du geste humain*. Cette action spécifique a entre autre débouché sur la tenue d'un atelier sur l'*Acquisition du geste humain par vision artificielle et applications* lors de la conférence RFIA 2004. De même, cette thématique s'intègre aux travaux développés dans le cadre du Gdr ISIS avec entre autre la tenue en janvier 2005 d'une journée sur l'Analyse de Visages, journée organisée dans le cadre du thème *E : Images, Modèles et Systèmes : Traitement, Analyse, Indexation.*
Les raisons de cet engouement reposent sur le fait qu'à l'heure actuelle, en traitement d'images, on ne peut plus se limiter à l'extraction d'informations bas-niveau n'ayant aucun sens. Au contraire, il faut se diriger vers des interprétations de plus haut niveau. Ceci est résumé dans [Bobick01] sous la forme suivante : la question fondamentale en traitement d'images n'est plus « *how are things* » mais elle est devenue « *what is happening ?* ».
Par ailleurs, il existe un potentiel important d'applications à ces recherches. On distingue :

- la simplification de la communication entre l'homme et la machine. L'idée est de rendre le dialogue homme/machine plus convivial. En plus d'un système de reconnaissance du langage, des informations issues d'une analyse vidéo sont tout à fait pertinentes (par exemple, développement d'une souris « optique »).
- des applications de réalité mixte pour lesquelles un utilisateur va interagir avec un monde virtuel. L'évolution du monde virtuel est alors dépendante du comportement, de l'état d'esprit et des activités de l'utilisateur.
- le développement de systèmes de « surveillance » intelligents. Par exemple, pour la surveillance de personnes âgées, il est pertinent de savoir si une personne est tombée et se trouve en situation de détresse. Un autre exemple réside dans la reconnaissance de comportements suspects (tels que des comportements agressifs) ou de comportement de panique dans les transports en commun, … Une autre possibilité concerne les systèmes de e-learning. Une WebCam suivie d'une analyse des images permettrait d'attirer l'attention du professeur sur les élèves soit en difficulté soit en manque de concentration.
- assistance en univers spécifique tels que les centrales nucléaires ou les aéroports pour lesquels l'analyse automatique de gestes est une aide précieuse.

Les travaux décrits dans la suite ne portent pas sur le thème des interactions homme/machine proprement dites. Ils se situent en amont et ils portent sur la problématique de savoir comment faire pour que la machine « comprenne le langage humain ». Le terme de langage ne fait pas

référence ici au message de parole. Il s'agit plutôt d'interpréter automatiquement un certain nombre d'informations liées à la communication non verbale : on distingue entre autre les gestes (de la main et de la tête), les postures, les expressions faciales. Le capteur envisagé est une caméra donc le support de l'information est l'image ou la vidéo.
Donnons une définition plus précise de toutes les actions considérées (voir Figure 1) :

- geste : *mouvement du corps, principalement de la main, des bras, de la tête, porteur ou non de signification* (définition du Petit Larousse)
- posture : *attitude particulière du corps* (définition du petit Larousse)
- expression faciale : *ensemble des signes du visage qui traduisent un sentiment, une émotion* (définition du Petit Larousse)

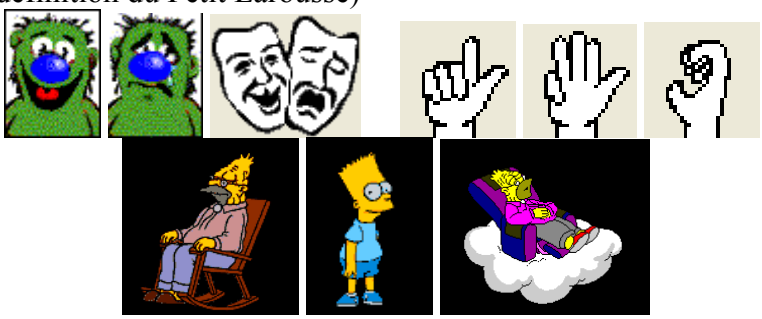

Figure 1 : exemples d'expressions faciales, de gestes de la main et de postures

Pour tous les «gestes» considérés, la méthodologie globale de reconnaissance ou d'interprétation est identique.

- Dans un premier temps, nous nous sommes attachés à extraire des informations de bas niveau (qui ne possèdent pas de signification sémantique) dans les séquences d'images. Plusieurs types d'informations bas niveau sont extraites : des contours, des régions et des informations de mouvement.
    - Pour l'extraction des contours de traits du visage tels que les yeux, la bouche et les sourcils, nous avons développé des algorithmes utilisant des modèles paramétriques de contours adaptés aux formes recherchées. Les paramètres des modèles de contours de type courbes de Bézier ou courbes cubiques sont estimés par des méthodes robustes de maximisation de flux de gradient de luminance et/ou de chrominance.
    - Afin d'extraire et de suivre les régions mobiles présentes dans une scène, nous avons développé une méthode de détection de mouvement basée sur une modélisation markovienne. Le suivi des régions extraites se fait par utilisation d'un filtre de Kalman à données éventuellement incomplètes dans le cas où la scène présente des occultations des objets mobiles.
    - Afin d'estimer automatiquement la nature et la direction des mouvements de tête d'une personne, nous avons développé une méthode fréquentielle efficace qui s'appuie sur la modélisation de certains des traitements qui se produisent au niveau du système visuel humain.
    - Nous avons développé une méthode de segmentation de la main qui s'appuie sur des informations de chrominance et sur une carte de transformée de distance et qui permet de tendre vers un modèle de main avec une probabilité de présence associée à chaque doigt.
- Dans un second temps, une phase de fusion de données est envisagée afin d'aboutir à une interprétation de haut niveau ayant une signification particulière. Bien que des méthodes de type HMM ou réseau bayesien aient été considérées à des fins de comparaison de performances de classification, la principale méthode de fusion de données utilisée est la théorie de l'évidence. En effet, vis à vis des problèmes de

classification de postures ou de classification d'expressions faciales, nous estimons que cette approche présente l'avantage de pouvoir considérer des mélanges d'expressions ou des mélanges de postures et d'autre part, elle permet la prise en compte explicite de données imprécises ce qui est important ici puisque toutes nos données résultent de méthodes de traitement d'images.

Dans la suite du manuscrit, nous présentons un système de reconnaissance d'expressions faciales (cf. section 5), un système d'analyse et d'interprétation de mouvements de la tête (cf. section 6), un système de reconnaissance de postures (cf. section 7) ainsi qu'un système de reconnaissance de gestes spécifiques (gestes du Langage Parlé Complété) (cf. section 8). La suite du manuscrit concerne l'ensemble des projets qui ont été associés à ces recherches (cf. section 9) et se termine par la présentation du projet de recherche (cf. section 10).

4.2 Positionnement dans le laboratoire

Je suis à l'origine du développement de la thématique d'analyse et d'interprétation de mouvements humains dans le laboratoire. Elle s'inscrit maintenant dans une thématique plus large qui est conséquente et qui porte sur la perception, l'humain et le multimédia. Par exemple, les travaux relatifs à l'interprétation des mouvements rigides de la tête s'appuient sur des modèles cognitifs de perception (cf. modélisation de la rétine humaine) développés au laboratoire depuis quelques années.
Dans le quadriennal qui vient de se terminer, les travaux relatifs à l'analyse des mouvements humains relevaient de deux groupes de recherche du laboratoire : le Groupe Objets, Traitement et Analyse (GOTA) et le groupe Signal Image et Communication (SIC). Par ailleurs, par le biais de la thèse Cifre de Thomas Burger portant sur le développement d'un système temps réel de reconnaissance des gestes du LPC, je collabore avec les chercheurs du groupe Circuits et Architectures.
Enfin, cette thématique est à l'origine de la mise en place au LIS d'une plate-forme vidéo d'interactions multimodales équipée du matériel approprié tant au niveau des capteurs (caméras mono et stéréo) qu'au niveau des machines de traitement.

5 Reconnaissance des expressions faciales

L'objectif de ces travaux est de définir un système dynamique et automatique de reconnaissance des expressions faciales. Pour ce faire, il faut résoudre trois tâches principalement [Pantic00a] : localiser le visage ; extraire de ce visage les informations utiles à la reconnaissance de l'expression faciale ; définir l'ensemble des expressions à reconnaître ainsi que le processus de classification associé. Si on considère le système humain comme le système de référence en terme de reconnaissance d'expressions, on constate que celui-ci s'appuie sur différentes modalités (audio, vidéo…) afin d'aboutir à la reconnaissance. Dans ce qui suit, le système de classification envisagé s'appuie principalement sur la modalité vidéo. Une ébauche de système multi-modal intégrant des informations issues de la vidéo et de l'audio est néanmoins présentée.

Le visage représente la partie du visage la plus expressive. Dans l'article [Mehrabian68], il a été démontré que, lors d'une communication face à face, le message transmis résulte du signal de parole, mais aussi de l'information transmise au travers de l'intonation ainsi que de celle transmise par l'intermédiaire des expressions faciales. D'un point de vue physiologique, une expression faciale résulte de la déformation de certaines parties du visage. Plus précisément, l'ensemble des déformations possibles au niveau du visage résultant de l'activité des différents muscles a été décrit par le système FACS (Facial Action Coding System) [Ekman78]. Ce système décrit chaque déformation faciale comme une combinaison de plusieurs actions élémentaires ou AU (Action Unit). Au total, il existe 44 AU différentes. La description d'une expression faciale quelconque à partir de ces unités d'actions étant complexe, beaucoup de travaux se limitent à la classification des expressions particulières que sont les émotions et qui provoquent effectivement des déformations au niveau du visage. C'est également ce que nous avons fait. Nous présentons dans ce mémoire un système axé sur la reconnaissance des six émotions universelles proposées par Ekman que sont *la joie, la peur, le dégoût, la tristesse, la colère* et *la surprise*. A ces six expressions, nous avons ajouté l'expression *neutre* ainsi qu'une catégorie d'expressions dites *inconnues* qui englobe toutes les expressions faciales différentes des six émotions universelles et de l'expression neutre.

Des études en psychologie ont mis en évidence le fait que l'être humain utilise le visage dans son ensemble pour reconnaître une expression mais aussi qu'il se focalise sur certains traits particuliers. Cette constatation est à l'origine de deux types d'approches pour la reconnaissance des expressions faciales : les approches s'appuyant sur un modèle global du visage et les approches travaillant à partir de l'analyse de certains traits caractéristiques. Tous nos travaux reposent sur l'hypothèse que **les expressions faciales peuvent être reconnues à partir des informations relatives aux déformations des traits permanents du visage que sont les yeux, les sourcils et la bouche**. Ce choix a été motivé par les expériences menées en psychologie pour lesquelles il a été démontré qu'il est possible à un être humain de reconnaître et d'interpréter une expression faciale en analysant les déformations au cours du temps des traits principaux du visage que sont les yeux, les sourcils et la bouche [Ekman99, Bassili78]. Nous avons pour notre part validé cette hypothèse en réalisant une expérimentation en psychologie [Sourd03]. Lors de cette expérience, menée en collaboration avec le Laboratoire de Psychologie Sociale de Grenoble et le CLIPS (équipe ARCADE), nous avons présenté à 60 sujets (30 hommes et 30 femmes) l'évolution temporelle des contours de la bouche, des yeux et des sourcils lors de la production d'une expression et nous leur avons demandé de reconnaître l'expression présente sur le visage sur la base de ces seuls contours. Un taux de reconnaissance de 60% a été obtenu. Notons que le taux de reconnaissance par un humain d'une expression sur un visage n'est jamais de 100% même si on présente l'ensemble

du visage à l'observateur et non uniquement les contours des traits qui se déforment. Selon les travaux de Bassili [Bassili78], ce taux de reconnaissance est en moyenne de 87%.
Par ailleurs, une difficulté majeure dans le domaine de la reconnaissance d'expressions faciales est l'obtention de base de données afin de valider les études proposées. L'obtention d'une base de données significative est un travail de très longue haleine. Tous les résultats proposés par la suite ont été validés sur des données acquises au laboratoire et sur certaines bases de données en accès libre. Une limitation demeure toutefois, toutes les données utilisées ont été produites par des acteurs professionnels ou non auxquels il a été spécifiquement demandé de simuler l'une des six émotions universelles. Nous décrirons les tentatives qui ont été faites ou qui sont en cours afin d'obtenir des expressions qualifiées de « naturelles » i.e. des séquences émotionnelles traduisant des émotions ayant été véritablement ressenties par un sujet.
La première étape à tout système d'analyse d'émotions concerne la localisation du visage dans l'image. De nombreux travaux ont été effectués à ce sujet. Le lecteur pourra se référer aux deux articles de synthèse [Hjelmäs01, Yang02] pour avoir une description détaillée des algorithmes existants. Nous n'avons pas abordé le problème de la localisation du visage. Dans nos travaux, nous avons utilisé le détecteur de visage développé par Viola et Jones [Viola04] dont le code est disponible sous forme d'une librairie nommée Machine Perception Toolbox [MPT] fonctionnant en C, en Matlab, sous Linux ou sous Windows.
Une fois le visage localisé (sous la forme d'une boite englobante), il s'agit d'en extraire les informations pertinentes qui conduiront à l'analyse des expressions. On distingue pour ce faire deux types d'approche :

- les approches qui considère le visage comme une entité globale dont les caractéristiques et les déformations sont apprises. Ces méthodes requièrent des phases significatives d'apprentissage. On peut citer pour exemple les algorithmes développés dans [Edwards98, Abboud04] qui utilisent une représentation du visage par modèle statistique d'apparence.
- les approches qui s'intéressent à des traits particuliers du visage afin d'en étudier les déformations. Les traits ou points caractéristiques pris en compte diffèrent d'une méthode à l'autre même si les yeux la bouche et les sourcils sont des caractéristiques prises en compte de manière dominante. La difficulté réside dans le développement d'algorithmes permettant l'extraction et/ou le suivi des traits sélectionnés et ce de manière automatique.

La dernière étape réside dans la phase de classification. Les classifieurs les plus couramment utilisés sont les réseaux de neurones, les classifieurs de type bayesien et les classifieurs à base de règles (résultant d'une expertise).
Pour plus de détails à propos des différents systèmes de reconnaissance d'expressions, le lecteur pourra se référer aux articles de synthèse [Fasel03, Pantic00b].
Nous proposons dans ce qui suit un algorithme de reconnaissance des six émotions universelles à partir des informations issues de l'analyse des déformations des contours des yeux, de la bouche et des sourcils. La classification est effectuée en utilisant la théorie de l'évidence.
Les travaux évoqués dans cette partie ont été menés dans le cadre des DEA de D. Lam, N. Mottin, A. Haddad, N. Eveno, Z. Hammal et Y. Lahaye puis des thèses de Nicolas Eveno soutenue en décembre 2003 et de Zakia Hammal dont la soutenance est prévue pour la fin de l'année 2005. Ils ont été détaillés dans les articles de revue suivants **[R_Eveno04, R_Hammal05a, R_Hammal05b]**. Un résumé des algorithmes développés est donné dans ce qui suit.

5.1 Extraction d'informations bas niveau : segmentation des contours des traits du visage (lèvres, yeux, sourcils)

Nous nous intéressons à l'extraction automatique des contours des traits permanents du visage à savoir : les yeux, les sourcils et les lèvres dans le but d'obtenir le squelette d'une émotion. La Figure 2 présente un exemple de squelette émotionnel afin de fixer les idées.

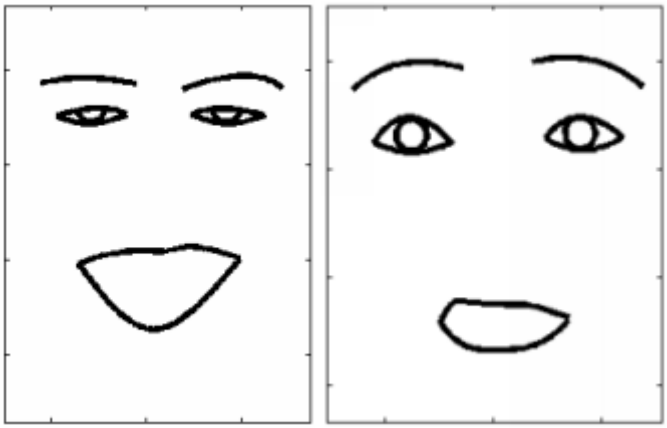

Figure 2 : squelette d'émotion : à gauche, la joie ; à droite, la surprise.

Pour chacun des traits considérés, un modèle paramétrique spécifique capable de rendre compte de toutes les déformations possibles est défini. Lors de la phase d'initialisation, des points caractéristiques du visage sont extraits (coins des yeux et de la bouche par exemple) et servent de points d'ancrage initiaux pour chacun des modèles. Dans la phase d'évolution, chaque modèle est déformé afin de coïncider au mieux avec les contours des traits présents sur le visage analysé. Cette déformation se fait par maximisation d'un flux de gradient (de luminance et/ou de chrominance) le long des contours définis par chaque courbe du modèle. La définition de modèles permet d'introduire naturellement une contrainte de régularisation sur les contours recherchés. Néanmoins, les modèles choisis restent suffisamment flexibles pour permettre une extraction réaliste des contours des yeux, des sourcils et de la bouche.

5.1.1 Pré-traitement :atténuation des variations d'éclairement

Dans une phase de pré-traitement, nous nous affranchissons des variations d'illumination en utilisant un filtrage adapté inspiré du comportement de la rétine [Beaudot94]. Ce filtre permet de réaliser un lissage local des variations d'éclairement. Une description plus détaillée de ce filtre est donnée à la section 6.
La Figure 3 présente l'effet de ce filtre d'une part sur une image de synthèse et d'autre part sur un visage éclairé latéralement. A l'issu du filtrage, les variations de luminance ont été fortement atténuées, le phénomène étant d'autant plus marqué pour l'image de synthèse.

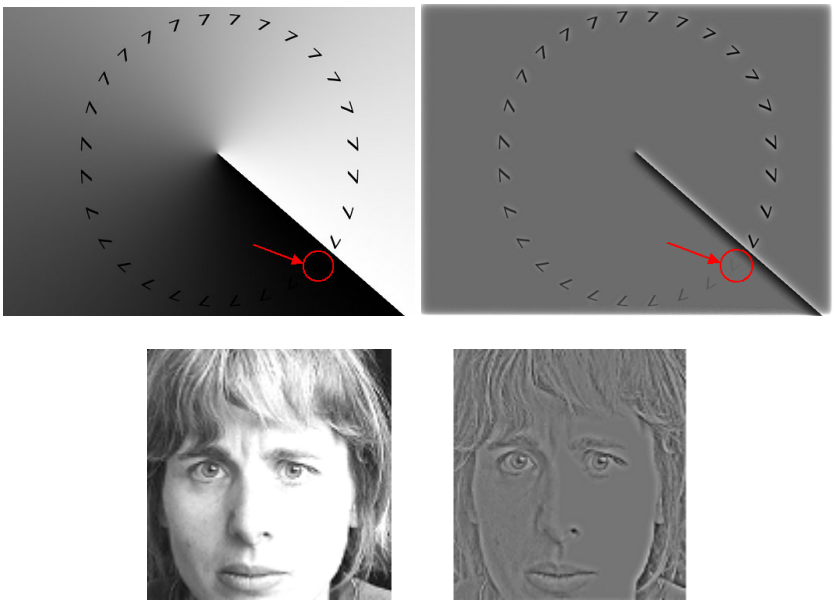

Figure 3 : à gauche, images à forte différence d'éclairement ; à droite, images en sortie du filtre de lissage des variations d'éclairement

5.1.2 Choix des modèles

Le choix d'un modèle adapté pour modéliser les lèvres est délicat car la forme des lèvres est très variable. Si le modèle choisi n'est pas bien adapté, le résultat de la segmentation ne sera pas de bonne qualité. Le modèle que nous proposons est composé de 5 courbes indépendantes, chacune d'entre elles décrivant une partie du contour labial. Entre Q_2 et Q_4, l'arc de Cupidon est décrit par une ligne brisée tandis que les autres portions du contour sont décrites par des courbes polynomiales cubiques γ_i (voir Figure 4-gauche). De plus, on impose à chaque cubique d'avoir une dérivée nulle au point Q_2, Q_4 ou Q_6. Par exemple, γ_1 (cubique entre Q_1 et Q_2) doit avoir une dérivée nulle en Q_2.

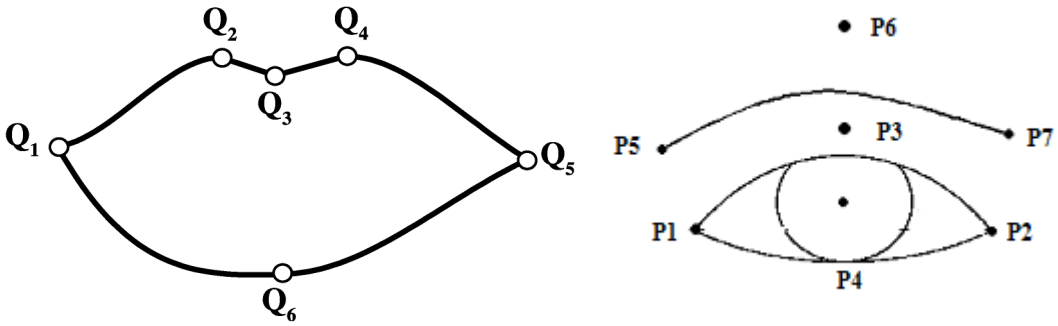

Figure 4 : modèles paramétriques proposés : à gauche, pour la bouche ;à droite, pour l'œil et le sourcil

Le modèle proposé pour les yeux et les sourcils est moins flexible que celui de la bouche car les yeux et les sourcils sont beaucoup moins déformables (cf. Figure 4-droite) :

- Pour chaque oeil : un cercle pour l'iris (éventuellement incomplet si l'œil est semi-ouvert) ; pour le contour inférieur, une parabole définie par trois points $\{P_1, P_2, P_4\}$; pour le contour supérieur, une courbe de Bézier à trois points de contrôle $\{P_1, P_2, P_3\}$; on se limite à une droite passant par P_1 et P_2 dans le cas d'un œil fermé.
- Pour les sourcils : une courbe de Bézier à trois points de contrôle $\{P_5, P_6, P_7\}$ pour le contour inférieur (on se limite à ce contour).

5.1.3 Extraction de points caractéristiques et initialisation des modèles

Afin d'initialiser les modèles sur les images à traiter, un ensemble de points caractéristiques est extrait. Il s'agit essentiellement des coins des yeux, de la bouche et des sourcils.

Cas des yeux

La première étape consiste en l'extraction du contour de chaque iris. Ce contour étant la frontière entre une zone sombre (l'iris) et une zone claire (le blanc de l'œil), il est recherché sous la forme d'un cercle constitué de points de gradient de luminance maximal. De plus, comme les yeux peuvent éventuellement être semi-ouverts, il arrive que le contour supérieur de l'iris soit occulté. Ainsi, on recherche le demi-cercle inférieur qui maximise le Flux de Gradient de luminance Normalisé :

$$FGN = \frac{1}{length\ SC} \sum_{p \in SC} \vec{\nabla} I(p).\vec{n}$$

Avec $I(p)$ luminance au point p, SC le demi-cercle cherché et $\vec{n}$ la normale au contour au point p. En effet, en chacun des points du cercle cherché, le gradient de luminance est normal au contour.
Connaissant la position des iris, il est alors possible d'extraire les coins des yeux par un processus de suivi de points de gradient localement maximal **[CN_Hammal03, CI_Hammal04]**. La Figure 5 donne une illustration de la méthode de détection des coins des yeux : en partant du point X_1 (point situé au niveau du point le plus bas du cercle détecté de l'iris et décalé de 2 pixels vers la gauche) un algorithme de suivi, vers la gauche, de points de gradient maximum conduit à la détection du coin C_1. La courbe joignant les points X_1 et C_1 est constituée d'un ensemble de points de gradient de luminance localement maximum. Le suivi de point s'arrête lorsque le gradient de luminance décroît fortement puisqu' alors on a atteint des points de peau situés au delà du coin de l'oeil. Un processus de suivi similaire permet de détecter le second coin C_2 en partant de X_2 (point situé au niveau du point le plus bas du cercle détecté de l'iris et décalé de 2 pixels vers la droite).

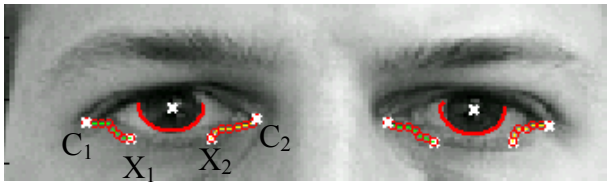

Figure 5 : détection des coins des yeux par suivi de points de gradient localement maximum

Cas des sourcils

Pour la détection des abscisses x_5 du coin intérieur P_5 et x_7 du coin extérieur P_7 (voir Figure 4), on recherche les abscisses des deux points pour lesquels il y a changement de signe ou annulation de la dérivée de la projection horizontale de l'image vidéo inverse. Pour la détection des ordonnées y_5 du coin intérieur P_5 et y_7 du coin extérieur P_7, on recherche l'abscisse du maximum de la projection verticale de l'image vidéo inverse **[CI_Hammal04, R_Hammal05a].**

Cas de la bouche

La détection de points caractéristiques sur la bouche en vue d'initialiser le modèle est plus complexe et elle se fait en utilisant conjointement une information hybride combinant la

luminance et la chrominance ainsi que la convergence d'un nouveau type de snake nommé « *jumping snake* ».
Afin de tenir compte de la couleur et de la luminance pour différentier les pixels de lèvre des pixels de peau, on utilise l'information hybride $\vec{R}_{top}(x,y)$, définie dans **[CI_Eveno02a]**. Elle est calculée de la manière suivante :

$$\vec{R}_{top}(x,y) = \vec{\nabla}\left[h_N(x,y) - I_N(x,y)\right]$$

où $h_N(x,y)$ et $I_N(x,y)$ sont respectivement la pseudo-teinte (définie dans l'espace RGB par $h(x,y) = \frac{R(x,y)}{G(x,y)+R(x,y)}$) et la luminance au pixel (x,y), normalisées entre 0 et 1. Cette information hybride permet de faire ressortir la frontière supérieure des lèvres beaucoup mieux que le gradient de luminance ou de pseudo-teinte seul **[CI_Eveno01]**.
Pour détecter des points caractéristiques du contour supérieur de la lèvre, nous définissons un nouveau type de contour actif que nous avons désigné sous le nom de *jumping snake* car sa convergence fait intervenir successivement des phases de croissance et des phases de sauts **[CI_Eveno03]**. Les trois points supérieurs sont situés sur le contour résultant de la convergence *du jumping snake* : Q_2 et Q_4 sont les points les plus hauts de part et d'autre de la ligne verticale de symétrie de la bouche. Q_3 est le point le plus bas du contour situé entre Q_2 et Q_4 (voir Figure 6).
Les points Q_6, Q_7 et Q_8 sont détectés par analyse de $\nabla_y(h)$, gradient 1D de la pseudo-teinte le long de l'axe vertical passant par Q_3 (voir Figure 6). Le maximum de $\nabla_y(h)$ au-dessous du contour supérieur donne la position de Q_7. Q_6 et Q_8 sont les minima de $\nabla_y(h)$ en-dessous et au-dessus de Q_7 respectivement. Ceci suppose que le visage est aligné sur la verticale et donc que les lèvres sont horizontales.

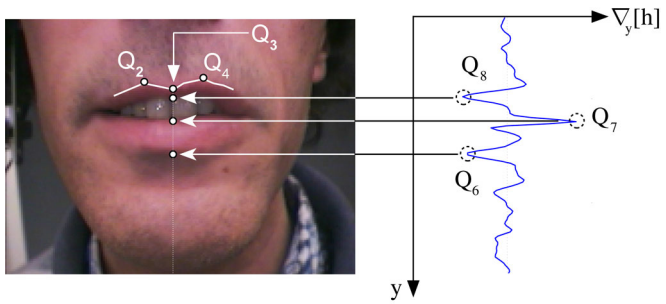

Figure 6 : les trois points du haut proviennent du jumping snake (ligne blanche). Q_6, Q_7 et Q_8 sont en dessous de Q_3 sur les extrêma de $\nabla_y[h]$.

5.1.4 Déformation des modèles initiaux

Ayant défini les modèles a priori et ayant extrait quelques points caractéristiques, il est possible de procéder à l'initialisation des contours recherchés. La Figure 7-gauche présente le résultat de l'initialisation pour les yeux et les sourcils et la Figure 7-droite présente le résultat de l'initialisation pour la bouche.

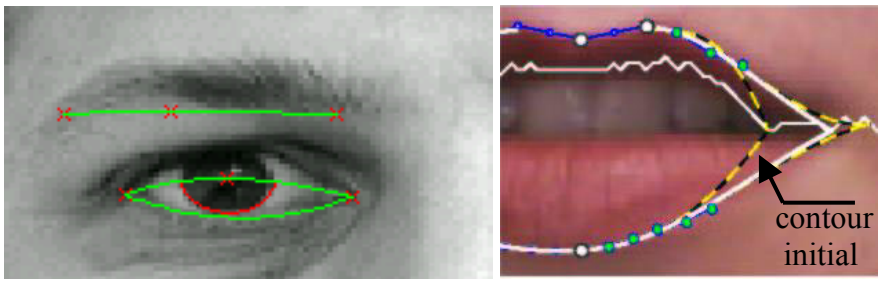

Figure 7 : à gauche, contours initiaux pour l'œil et le sourcil ; à droite, contour initial pour la bouche

Chacun des contours initiaux est déformé de manière à coïncider avec le contour présent sur l'image. Cette déformation nécessite la maximisation du flux de gradient de luminance pour les yeux et les sourcils **[R_Hammal05a, R_Hammal05b]** et la maximisation du flux de gradient hybride pour la bouche **[CI_Eveno03, CN_Hammal03]**. En ce qui concerne le contour des lèvres, nous avons proposé un algorithme d'optimisation qui recherche en même temps le meilleur contour et les meilleurs coins de la bouche.

5.1.5 Résultats

La Figure 8 présente quelques résultats types obtenus à l'issu de la phase de segmentation des traits du visage. Ces résultats permettent d'une part de juger de la qualité de la segmentation et d'autre part, ils permettent de mettre en évidence la pertinence des modèles choisis qui sont suffisamment flexibles pour rendre compte de formes différentes.

Figure 8 : résultats de segmentation

5.2 Système de reconnaissance d'expressions faciales basé sur la vidéo

Les contours obtenus à la suite de l'étape de segmentation étant suffisamment précis et réalistes, ils sont utilisés comme information bas niveau pour le système de reconnaissance d'expressions faciales. Ces travaux ont conduit aux publications [**CN_Hammal04a, CI_Hammal05c, CI_Hammal05e**].

Dans le système de reconnaissance d'expressions faciales développé, nous proposons d'utiliser la théorie de l'évidence [Smets98, Dempster68, Shafer76] qui semble tout à fait

adaptée à ce problème de classification. En effet, contrairement à d'autres méthodes, la théorie de l'évidence permet de tenir compte de données imprécises ce qui ne manquera pas d'arriver étant donné que toutes nos mesures résultent d'algorithmes de segmentation de contours. Par ailleurs, la théorie de l'évidence permet de tenir compte du caractère non totalement binaire des expressions recherchées. En effet, une expression n'est pas toujours exprimée de manière unique : certaines expressions peuvent être un mélange de plusieurs émotions de base. Par exemple, on constate souvent un mélange des deux émotions *peur* et *surprise*. Le formalisme proposé par la théorie de l'évidence permet de rendre compte d'expressions mélangées.

5.2.1 Les mesures

Afin de réaliser la reconnaissance des expressions faciales, les mesures prises en compte sont les cinq distances définies sur la Figure 9.

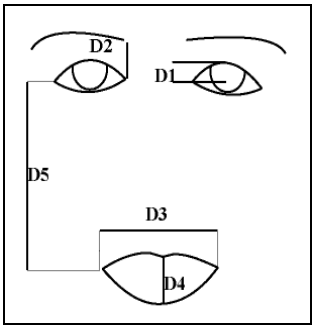

Figure 9 : distances caractéristiques définies sur le squelette émotionnel

On distingue :

- D_1 qui mesure le degré d'ouverture des yeux
- D_2 qui mesure l'écartement entre les yeux et les sourcils
- D_3 et D_4 qui caractérisent l'ouverture de la bouche
- D_5 qui mesure la distance entre les coins de la bouche et les coins des yeux.

Lors de la production d'une émotion, ces distances évoluent puisque les traits du visage se déforment. La Figure 10 présente l'évolution de deux de ces distances pour plusieurs sujets différents simulant la même émotion. On constate qu'en cas de surprise, la distance D_2 augmente pour tous les sujets : les yeux s'ouvrent (cf. écarquillement). En cas de joie, la distance D_5 diminue pour tous les sujets : les coins de la bouche se rapprochent des sourcils (cf. grand sourire).

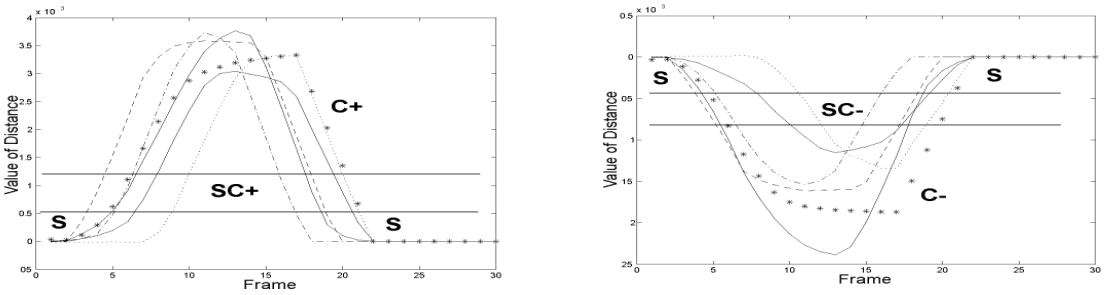

Figure 10 : à gauche, évolution de D_2 en cas de *surprise* ; à droite, évolution de D_5 en cas de *joie*

Soit D_i^{ref}, la valeur de chacune de ces 5 distances pour l'expression neutre, les D_i^{ref} sont considérées comme des distances de référence. Nous proposons d'associer l'un des trois états symboliques suivants à chacune des valeurs des distances D_i :

- état S pour lequel la distance *Di* est du même ordre de grandeur que sa valeur pour l'expression neutre ;
- état C^+ pour lequel la distance *Di* est plus grande que sa valeur pour l'expression neutre ;
- état C^- pour lequel la distance *Di* est plus petite que sa valeur pour l'expression neutre.

La délimitation des différents états est matérialisée par des lignes horizontales sur la Figure 10. On constate de plus que des zones de doute pour lesquelles deux états sont possibles sont introduites (états *SC*+ et *SC*-) afin d'éviter une prise de décision trop brutale.
Suite à une expertise sur notre base de séquences vidéo d'expressions, nous avons défini pour chaque expression une combinaison spécifique d'états (cf. Tableau 1)

	D_1	D_2	D_3	D_4	D_5
Joie E_1	C^-	S / C^-	C^+	C^+	C^-
Surprise E_2	C^+	C^+	C^-	C^+	C^+
Dégoût E_3	C^-	C^-	S / C^+	C^+	S / C^-
Colère E_4	C^+	C^-	S	S / C^-	S
Tristesse E_5	C^-	C^+	S	S	S
Peur E_6	S / C^+	S / C^+	S / C^-	S / C^+	S
Neutre E_7	S	S	S	S	S

Tableau 1 : ensemble des états symboliques associé à chaque expression

D'après le Tableau 1, on constate que dans le cas de la *surprise*, D_3 diminue et toutes les autres distances augmentent. En effet, en cas de *surprise*, on a tendance à ouvrir la bouche et à rester bouche bée (D_4 augmente, D_3 diminue) et à écarquiller les yeux et les sourcils (D_1,D_2 augmentent). On remarque par ailleurs que pour certaines expressions, une distance peut être associée à deux états (voir le cas de la *peur* par exemple). Ceci provient de la variabilité entre les individus lors de la production de certaines émotions. Il résulte de cette variabilité que dans certains cas deux émotions ne pourront pas être différenciées (par exemple, possibilité de confusion entre la *joie* et le *dégoût* ou entre la *peur* et le *neutre*) à partir des seules distances D_1 à D_5. Des informations supplémentaires doivent être considérées.
La table de règles ainsi définie est compatible avec la description des expressions faciales fournie dans la norme MPEG-4 et reportée dans [Malciu01].

5.2.2 La théorie de l'évidence appliquée à la reconnaissance d'expressions

Initialement introduite par Dempster [Dempster68], la théorie de l'évidence fut reprise par Shafer [Shafer76]. Smets a développé cette théorie qu'il appelle TBM (*Transferable Belief Model*) [Smets98]. Cette théorie peut être considérée comme une généralisation de la théorie des probabilités. Elle nécessite la définition d'un ensemble de définition Ω composé de N hypothèses H_i exclusives et exhaustives.

Dans cette théorie, le raisonnement porte sur le cadre de discernement 2^Ω qui est l'ensemble des 2^N sous-ensembles A de Ω.

Pour exprimer le degré de confiance d'une source d'information pour chaque élément A de 2^{Ω}, on lui associe une masse d'évidence élémentaire $m(A)$ qui indique toute la confiance que l'on peut avoir dans cette proposition sans pour autant privilégier aucune des classes qui la composent. La fonction *m* est définie par :

$$m: \quad 2^{\Omega} \to [0,1]$$
$$A \mapsto m(A)$$

avec : $\sum m(A) = 1$

Pour l'application considérée ici, les hypothèses E_i de Ω qui correspondent aux expressions faciales sont au nombre de 7 : sourire (E_1), surprise (E_2), dégoût (E_3), colère (E_4), tristesse (E_5), peur (E_6) et neutre (E_7).

Le but est de déterminer à chaque instant l'état associé à chaque mesure D_i. Pour ce faire, on définit la distribution de masse d'évidence m_{Di} qui indique pour chacun des états possibles $\{C^+, C^- S, SC^+, SC^-\}$ le degré de confiance que l'on a que la distance D_i soit dans cet état donné. On réalise ainsi une conversion numérique/ symbolique, qui associe à chaque valeur de D_i l'un des symboles $\{C^+, C^-, S, SC^+, SC^-\}$ avec une masse de croyance associée. Cette masse de croyance traduit la confiance que l'on a que la mesure D_i soit dans l'un des 5 états possibles.

Pour réaliser cette conversion, toute la difficulté réside dans la définition du modèle à associer à chaque distance. Le modèle choisi est décrit sur la Figure 11.

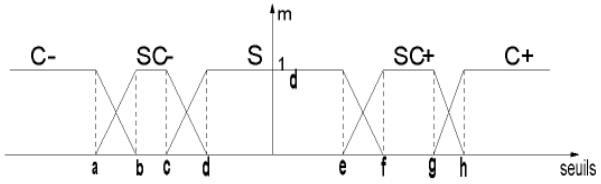

Figure 11 : modèle choisi

m représente la masse de croyance associée à chaque état possible et les seuils *(a…h)* sont les valeurs limites de D_i correspondant à chaque état ou sous ensemble d'états.

Dans le cas général, il faudrait pouvoir définir ces seuils par apprentissage statistique quand on dispose d'un grand nombre d'échantillons. Dans notre cas, cela a été fait par expertise sur l'ensemble des images de la base d'apprentissage.

A partir du Tableau 1, on peut en déduire une base de règles pour les 5 distances et les 7 expressions considérées. Par exemple, si on considère la distance D_1, on obtient le Tableau 2 pour lequel la valeur « 1 » est donnée lorsque la distance atteint l'état considéré et la valeur « 0 » est donnée dans le cas contraire.

		E_1	E_2	E_3	E_4	E_5	E_6	E_7
D_1	C^+	0	1	0	1	0	0/1	0
	C^-	1	0	1	0	1	0	0
	S	0	0	0	0	0	0/1	1

Tableau 2 : table de conditions binaires pour la distance D_1

A partir du Tableau 2, on peut en déduire les masses de croyance associée à chaque expression ou combinaison d'expressions et issues de l'information portée par la distance D_1. On obtient alors :

$$\boldsymbol{D_1} \quad m_{D1}(E_2 \cup E_4 \cup E_6) = m_{D1}(C^+)$$

$$m_{D1}(E_1 \cup E_3 \cup E_5) = m_{D1}(C^-)$$
$$m_{D1}(E_6 \cup E_7) = m_{D1}(S)$$
$$m_{D1}(E_2 \cup E_4 \cup E_6 \cup E_7) = m_{D1}(S \cup C^+)$$
$$m_{D1}(E_1 \cup E_3 \cup E_5 \cup E_6 \cup E_7) = m_{D1}(S \cup C^-)$$

En procédant de même pour l'ensemble des 5 mesures, on obtient un ensemble de masse de croyances élémentaires qu'il faut alors fusionner afin d'obtenir la masse de croyance finale associée à chaque expression ou ensemble d'expressions. Cette fusion est obtenue en utilisant la règle de combinaison conjonctive appelée somme orthogonale. Dans le cas de 2 distances D_1 , D_2 la somme orthogonale m est définie de la manière suivante:

$$m = m_{D_1} \oplus m_{D_2}$$
$$m(A) = \sum_{B \cap C = A} m_{D_1}(B).m_{D_2}(C)$$

où *A, B* et *C* sont des expressions ou sous ensembles d'expressions. La combinaison a pour effet d'affecter la masse d'évidence à des propositions dont le nombre d'éléments est plus faible que celui des propositions initiales.

Afin de pouvoir gérer les cas de conflit, on ajoute une expression E_8 dite *inconnue* ou de *rejet* représentant toutes les expressions qui ne correspondent à aucune des descriptions du Tableau 1.

La prise de décision consiste à faire un choix entre les différentes hypothèses E_i et leurs combinaisons possibles. Le choix implique obligatoirement une prise de risque, sauf si le résultat de la combinaison est sure : $m(E_i) = 1$. On cherche à optimiser un critère et en l'occurrence dans ces travaux, nous faisons le choix de l'expression qui recueille la masse de croyance la plus élevée.

5.2.3 Résultats

Nous présentons ici quelques résultats de classification. Les seules expressions ayant été testées sont la *joie*, la *surprise*, le *dégoût* et le *neutre* car ce sont les seules expressions pour lesquelles nous avons suffisamment de données. En effet, dans la base de données que nous avons construite en demandant à des sujets non acteurs de simuler une expression, nous avons rejeté les expressions de *colère*, *tristesse* et *peur* car nous nous sommes aperçus que ces expressions sont très difficiles à simuler. Par ailleurs, parmi les bases de données existantes, seules celles de Kanade-Cohn [Cohn] et de Cotrell [Dailey01] ont pu être exploitées. Toutes les bases existantes d'expressions statiques (une seule image) et/ou d'images en niveau de gris n'ont pas pu être utilisées vue la méthode de classification que nous avons développée.

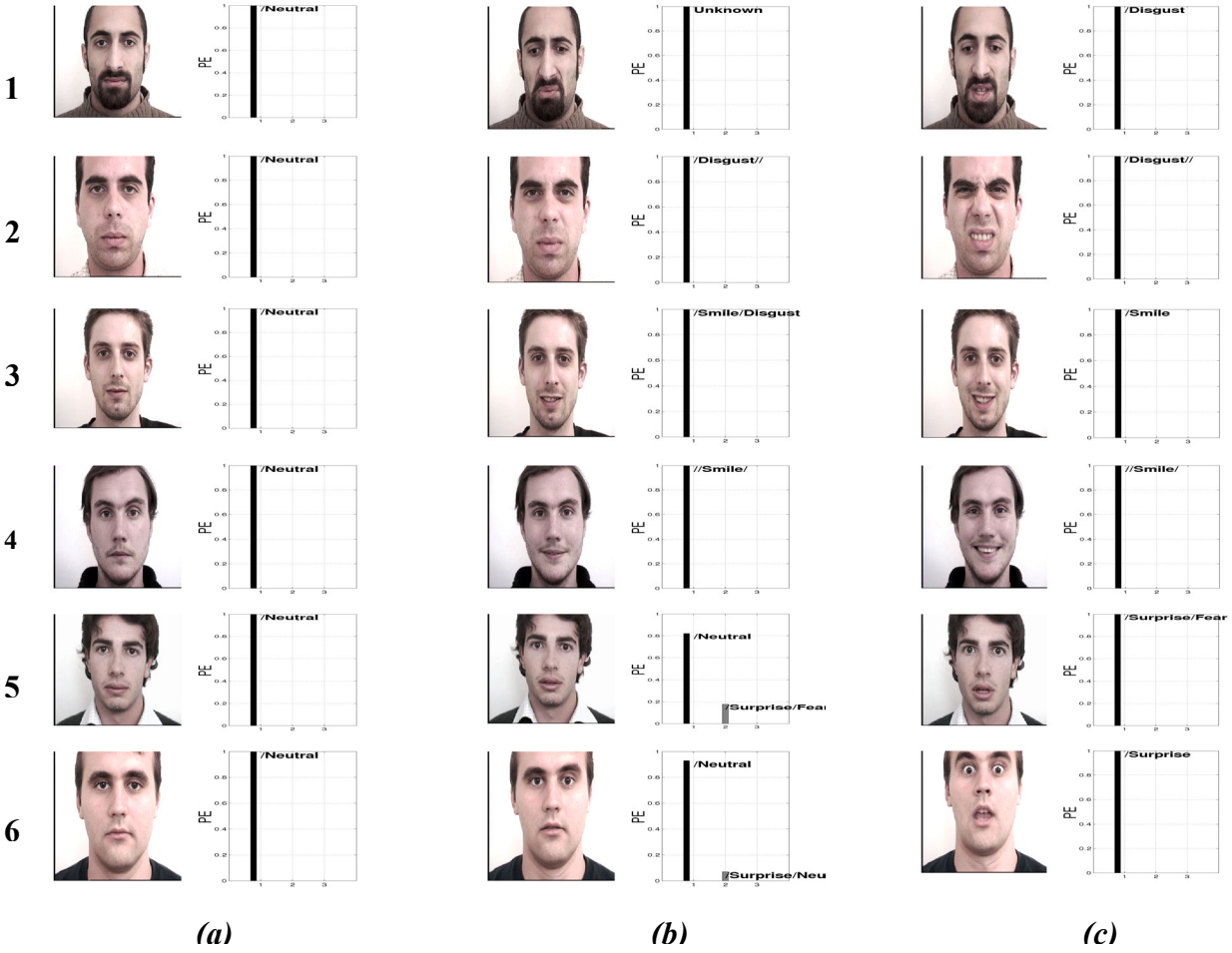

Figure 12 : exemples de classification : a) expression neutre ; b) expression intermédiaire ; c) apex de l'expression. Pour chaque cas sont donnés l'image ainsi qu' un graphe des masses d'évidence associées à chacune des expressions ou combinaisons d'expressions possibles (seules les masses d'évidence non nulles sont représentées)

La Figure 12 présente des exemples de classification. Les lignes 1 et 2 montrent l'aptitude du système à reconnaître une même expression (le *dégoût*) mais avec une intensité différente. La Figure 12-b-1 montre un cas pour lequel le choix final s'est porté sur l'expression de *rejet* ou expression *inconnue*. Cette expression est effectivement une expression de transition produite lors du passage de l'état *neutre* vers l'état de *dégoût*. Les lignes 3 et 4 présentent des résultats pour l'expression de *joie* et les lignes 5 et 6 montrent des résultats pour l'expression de *surprise*. La ligne 5 met en évidence la difficulté du système à séparer la *surprise* et la *peur*. Notons que même pour un classifieur humain, cette distinction s'avère difficile. Cependant, dans ce cas là, le système est sûr à 100% que l'expression présente sur le visage est soit de la *peur* soit de la *surprise* sans être capable de trancher entre les deux.
Les Tableau 3 et Tableau 4 donnent des résultats obtenus à l'issu du traitement de l'ensemble des données disponibles :

- base de données « maison » : 630 images de 7 sujets différents et 4 expressions ;
- base de Cohn-Kanade : 144 images et 4 expressions ;
- base de Cotrell : 24 images et 4 expressions

Pour les deux dernières bases de données, nous ne disposons que d'une image à l'état neutre et d'une image avec l'expression considérée. Il ne s'agit donc pas de séquences d'images. Ces

données nous permettent néanmoins de tester les performances de la méthode de classification statique proposée.

Syst\Exp	E_1	E_2	E_3	E_7
E_1 joie	76,36%	0	9,48	3%
E_2 surprise	0	12%	0	0
E_3 dégoût	0	0	43,10%	2%
$E_1 \cup E_3$	10,90%	0	8,62%	0
$E_2 \cup E_6$	0	72,44%	0	0
E_7 neutre	6,66%	0,78%	15,51	88%
E_8 inconnue	6,06%	11,8%	12,06%	0
autre	0.02%	2.08%	11.32%	7%
Total	**87,26%**	**84,44%**	**51,72%**	**88%**

Tableau 3 : matrice de confusion sur les données de la base « maison »

Afin d'établir les taux de classification totaux (cf. dernière ligne du Tableau 3), nous considérons comme bonne classification les % associés aux combinaisons d'expressions (pour peu que la combinaison contienne la « bonne » expression) car nous avons vu que le système de règles sur lequel repose la classification (cf. Tableau 1) peut engendrer des confusions. Seules des informations supplémentaires permettraient de lever l'ambiguïté.
L'étude du Tableau 3 montre que le système a plus de difficultés à reconnaître le *dégoût* que les autres expressions. Cette tendance, bien que plus faible, se retrouve sur les résultas du Tableau 4. Nous pensons que ceci est lié au fait qu'il est plus difficile de simuler le *dégoût* que les autres expressions considérées. Le fait que les taux de classification soient meilleurs sur le Tableau 4 s'explique par le fait que dans le cas des bases de Cohn-Kanade et de Cotrell, les images traitées sont uniquement des images dans lesquelles l'expression considérée est à son apogée ce qui n'est pas forcément le cas pour les images de la base « maison ».

Syst\Exp	E_1	E_2	E_3	E_7	E_1	E_2	E_3
E_1joy	64,51%	0	0	0	62,50%	0	0
E_2surprise	0	16%	0	0	0	25%	0
E_3disgust	0	0	52,94%	0	0	0	75%
$E_1 \cup E_3$	32,25%	0	47,05%	0	37,50%	0	0
$E_2 \cup E_6$	0	84%	0	0	0	75%	0
E_7neutral	0	0	0	0	0	0	0
E_8unknow	3,22%	0	0	0	0	0	25%
others	0	0	0.01%	0	0	0	0
Total	**100%**	**100%**	**99,99%**	**100%**	**100%**	**100%**	**75%**

Tableau 4 : matrices de confusion pour la base de Cohn-Kanade (à gauche de la colonne E_7) et pour la base de Cotrell (à droite de colonne E_7). La colonne E_7 est la même pour les deux bases de données.

5.3 Comparaison avec le classifieur idéal : le système de reconnaissance humain

Dans l'article [Pantic00b], 12 critères globaux, indépendants de toute application, sont associés au classifieur d'expressions supposé idéal, à savoir le classifieur humain. Afin de se faire une idée précise des performances du système proposé par rapport au classifieur humain, nous allons faire la liste de ces 12 critères et voir en quelle mesure, notre système satisfait ou non à chacun de ces critères. Par convention, les critères mis en gras sont ceux qui sont satisfaits par notre système.

1. ***Automatic facial image acquisition*** : dans notre système, les images sont acquises automatiquement par une WebCam ou par une caméra numérique SONY.
2. *Subjects of any age and ethnicity* :
    - problème de l'âge : avec des personnes âgées, il y a des rides supplémentaires permanentes sur le visage ce qui n'est pas pris en compte pas la plupart des systèmes qui procèdent à partir de l'apprentissage des déformations du visage

dans son ensemble. Dans notre système, la base de règles décrivant chacune des expressions à reconnaître reste la même quelque soit l'âge de la personne.

- problème de la race : pour des raisons aisément compréhensibles, la plus grande partie de nos tests a été effectuée sur des individus de race européenne. Une limitation concerne l'algorithme de détection des sourcils qui suppose implicitement que les sourcils sont plus sombres que la peau ce qui n'est pas le cas pour des individus très blonds. Par ailleurs, nous savons d'ores et déjà que l'algorithme utilisé pour localiser le visage ne fonctionne pas pour des individus à peau noire. De même, étant donnée que l'extraction de certains contours du visage repose sur la détection de gradient de luminance et/ou de chrominance, ces gradients sont moins prononcés pour les individus de race noire. Peu de tests ont pu être menés à ce sujet mais il est fort probable que les algorithmes d'extraction des données et donc par voie de conséquence l'algorithme de reconnaissance ne soient pas robustes dans ce cas là.

3. ***Deals with variations in lighting*** : le filtrage préalable des images à analyser par un filtre issu de la modélisation du comportement de la rétine humaine permet d'atténuer les variations d'illumination (cf. 5.1.1). Notre algorithme est donc robuste aux variations d'éclairement.
4. ***Deals with partially occluded faces*** : le système proposé ne fonctionne pas si les traits permanents considérés sont occultés. En effet, tout repose sur le fait que les contours des yeux, de la bouche et des sourcils sont visibles. Une remarque néanmoins : le système est capable de traiter le cas de personnes portant des lunettes de vue.
5. ***No special markers/make up required*** : nous travaillons directement sur les images acquises par la caméra, aucun artifice préalable sur la personne à filmer n'étant nécessaire.
6. *Deals with rigid head motion*: le système proposé suppose que le visage de la personne est proche de la verticale et qu'il est quasi immobile. Ceci constitue une limite pénalisante du système à l'heure actuelle.
7. ***Automatic face detection*** : bien que présentant des failles (personnes à peau noire, visages penchés), la détection du visage par l'algorithme de la librairie MPT se fait de manière automatique. D'autres détecteurs de visages sont à l'étude et en particulier celui proposé par Intel [OpenCV].
8. ***Automatic facial expression data extraction*** : l'extraction des contours du visage est entièrement automatique.
9. ***Deals with inaccurate facial expression data*** : l'utilisation de la théorie de l'évidence pour la classification permet justement de tenir compte de données imprécises.
10. ***Automatic facial expression classification*** : la classification est entièrement automatique, elle ne requiert aucune intervention manuelle.
11. *Distinguishes all possible expressions* : le système est limité aux six émotions universelles et à l'émotion neutre. Une caractéristique intéressante est qu'il est possible de détecter la présence sur un visage d'une expression autre que les 7 considérées même si la qualification précise de cette expression n'est pas faite. Le système proposé pourrait s'étendre à la reconnaissance d'autres expressions pour peu que l'on soit capable d'enrichir notre base de règles et de description d'une expression à partir des déformations des distances *Di* considérées.
12. *Deals with unilateral facial changes* : le système proposé suppose que les mêmes déformations se produisent sur les deux côtés du visage.

Notre système satisfait 8 critères sur les 12 proposés ce qui en fait un système relativement performant. Parmi les critères non satisfaits, le plus limitant est le critère de quasi mobilité du visage ce qui n'est pas réaliste pour bien des applications.

5.4 Vers un système multi-modal : prise en compte de l'information audio.

Nous avons vu au paragraphe 5.2, que notre système se trouve parfois devant des situations où il lui est impossible de choisir entre plusieurs expressions. Afin de lever ces ambiguïtés, nous proposons d'ajouter des informations supplémentaires issues du canal audio. En effet, il a été mis en évidence que les émotions peuvent se retrouver dans le signal de parole [Justin03, Scherer03, Schröder03].

Nous présentons un système de classification d'expressions vocales. Ce travail a été effectué dans le cadre de la thèse de Zakia Hammal et en collaboration avec l'équipe de Thierry Dutoit du laboratoire de traitement du signal de l'université de Mons, Belgique (ces recherches ont été financées en partie par le réseau d'excellence Similar). Trois publications ont résulté de cette collaboration **[CI_Hammal05a, CI_Hammal05b, CN_Hammal05].**

Dans un premier temps, une classification en 5 classes d'expressions (*Neutre*, *Surprise*, *Joie*, *Tristesse* et *Colère*) utilisant un ensemble de caractéristiques statistiques acoustiques (telles que le débit, la fréquence fondamentale, l'énergie, la proportion de HF par rapport aux BF…) a été envisagée sur la base d'expressions vocales DES [DES]. Des confusions entre groupes d'expressions ont été obtenues. Il s'avère que ce sont les mêmes confusions que celles obtenues lors d'une classification par un humain. Contrairement aux travaux classiques qui s'efforcent alors de trouver des caractéristiques supplémentaires permettant de dissocier les expressions les plus semblables, nous avons proposé de nous limiter à deux classes d'expressions seulement : la classe des voix *Agitées* regroupant la *Joie*, la *Surprise* et la *Colère* et la classe des voix *Calmes* regroupant le *Neutre* et la *Tristesse*. Cette classification présente l'avantage d'être plus conforme à la réalité. En effet, il n'est pas rare en situation réelle d'obtenir un mélange d'expressions dans la voix.

Pour valider le bien-fondé de ces deux nouvelles classes, plusieurs méthodes de classification ont été testées. On distingue un classifieur Bayesien, une classification par Analyse Linéaire Discriminante, un classifieur aux K plus proches voisins (KNN) et un classifieur à support vecteur machine (GSVM). Pour les deux classes considérées, les meilleurs taux de classification ont été obtenus avec le classifieur GSVM avec un taux de classification de 89.74% pour les voix *Agitées* et 86.54 % pour les voix *Calmes*.

Nous disposons donc maintenant d'un système de classification en 7 classes s'appuyant sur des informations vidéo et d'un système de classification en 2 classes s'appuyant sur des informations audio. Afin d'étudier la complémentarité éventuelle de ces deux canaux d'informations, il faut maintenant définir un système de classification multi-modal. Ce travail est en cours d'étude. Nous nous heurtons ici à un problème difficile : nous ne disposons pas de données audio-vidéo. Toutes les recherches entreprises pour récupérer sur le Web des données ont été vaines. Quant à la mise en place d'une campagne d'acquisition, c'est un problème long et compliqué. Une alternative serait d'extraire des séquences audio-vidéo sur des DVD.

5.5 Vers un système de reconnaissance d'expressions naturelles ?

Tous les résultats de classification obtenus jusqu'à aujourd'hui ont résulté de l'analyse de séquences audio ou vidéo pour lesquelles les émotions considérées ont été simulées soit par des acteurs soit par des non acteurs. Aucune des données utilisées ne présente des émotions ressenties, émotions que l'on qualifie ici de « naturelles » par opposition aux émotions

simulées par des acteurs. Afin d'aller plus loin dans la validation du système de classification proposé, en particulier celui travaillant sur la modalité vidéo, il faudrait pouvoir disposer de séquences d'émotions « naturelles ». Une ébauche de solution a été envisagée suite à une collaboration avec le Laboratoire de Psychologie Sociale de Grenoble et le CLIPS (équipe ARCADE). Pour permettre à des sujets en situation de travail devant un écran d'exprimer certaines émotions, il leur a été demandé de réaliser quatre tâches attentionnelles destinées à susciter ces émotions caractéristiques à savoir :

- *plaisante et extériorisée* (émotion de gaieté) : présentation d'une série de 12 images à caractère humoristique.
- *plaisante et intériorisée* (émotion d'intérêt) : les personnes doivent rechercher sur un site web la fréquence de leur patronyme et de leur prénom depuis le début du siècle jusqu'à aujourd'hui.
- *déplaisante et extériorisée* (émotion d'énervement) : jeu de « mikado » sur écran informatique à l'aide d'une souris dont les mouvements ne peuvent être que partiellement et aléatoirement maîtrisés entravant ainsi les performances du joueur.
- *déplaisante et intériorisée* (émotion d'anxiété) : il s'agit d'un « test de QI » repris d'un site web et modifié pour que le sujet ne puisse jamais répondre convenablement. De plus, le sujet est informé que ce test est en moyenne bien réussi par des étudiants de son niveau pour le situer dans une situation de compétition sociale générant du stress.

A ces quatre tâches, il a été ajouté une *tâche contrôle* de simple attention (lecture d'une notice explicative d'un jeu de société) correspondant à une émotion « *neutre* ». Toutes ces tâches ont été testées au préalable sur une vingtaine de personnes pour s'assurer qu'elles suscitent bien l'émotion recherchée.

Pour l'acquisition des séquences vidéo, la personne doit réaliser différentes tâches-émotions s'enchaînant assez rapidement (3-4 minutes par tâche) pour permettre de changer de registre émotionnel en passant d'une tâche à l'autre. Le but est de mettre la personne face à un protocole de tâches qui permet de susciter différentes émotions afin de pouvoir saisir et enregistrer les différents mouvements faciaux qui se succèdent. Cette passation a été effectuée auprès de 50 personnes environ.

Ensuite, des juges ont évalué individuellement l'expressivité de chaque « film-émotion » pour chaque tâche-sujet à l'aide d'échelle (de 1 « peu expressif » à 5 « très expressif » pour chaque émotion considérée). Suite à l'analyse des scores d'expressivité, les deux visages les plus expressifs pour chaque émotion ont été sélectionnés.

On constate que l'acquisition de séquences vidéo d'expressions réelles est un travail lourd et long à mettre en œuvre. Les séquences vidéo acquises lors de ces expérimentations n'ont pu être que très partiellement exploitées pour les raisons suivantes :

- il n'est pas simple de proposer des protocoles expérimentaux permettant de susciter les 6 émotions universelles proposées par Ekman. Lors des expériences proposées ici, 5 émotions ont été visées et elles sont parfois différentes de celles d'Ekman : *gaieté*, *intérêt*, *énervement*, *anxiété, neutre*. Il faut donc revoir les modèles proposés afin de construire un système compatible avec les expressions envisagées ici, ce qui est tout à fait possible mais qui n'a pas encore été mis en oeuvre.
- des mouvements de la tête ou des mains étant amenées à occulter partiellement le visage ont été fréquemment constatés ce qui limite le nombre de mesures disponibles pour la classification.
- sur les séquences filmées, les traits du visage ne sont pas toujours suffisamment précis (tout dépend de l'éloignement par rapport à la caméra) pour que nous puissions appliquer notre système de classification vidéo. En revanche, même lorsque le visage est un peu éloigné, un observateur humain est capable de reconnaître l'expression ce qui laisse à penser qu'il s'appuie sur d'autres informations que la déformation des

traits du visage (lorsque ceux ci ne sont pas suffisamment précis). Par exemple, il utilise des informations liées aux mouvements et à la position de la tête (tête basse quand on est honteux).

Ce sont ces deux dernières constatations qui sont à l'origine des travaux décrits dans la section 6.

5.6 Ce qu'il reste à faire

Les travaux sur l'analyse des expressions faciales ont débuté dans le cadre de la thèse de Zakia Hammal. Bien que profitant des résultats acquis en extraction de contours de lèvres lors de la thèse de Nicolas Eveno, ces travaux sont donc récents. Je vois le système développé jusqu'ici comme le point de départ de nombreux travaux à venir sur le sujet. En particulier, il s'agira :

- de modifier le système afin qu'il puisse tenir compte non pas des déformations présentes sur le visage dans une image à un instant donné mais afin qu'il puisse s'appuyer sur l'analyse de ces déformations au cours du temps. L'objectif est donc de développer un système dynamique de reconnaissance d'expressions. Cet aspect a été ébauché dans le cadre de la thèse de Zakia Hammal mais il devra être poursuivi. Il pose le problème de savoir comment utiliser la théorie de l'évidence dans un cadre dynamique. Peu de travaux ont été faits à ce sujet jusqu'à présent.
- de voir comment développer un système de reconnaissance d'expressions qui puisse tenir compte des mouvements de tête.
- de réfléchir à un système multi-modal s'appuyant sur des informations audio et vidéo. Une analyse des émotions dans la voie a été proposée. L'objectif sera de voir comment fusionner les informations provenant de l'audio et de la vidéo. Chaque information est elle indépendante ? Comment définir un ensemble d'expressions à reconnaître qui soit compatible avec les deux types d'information (pour le moment, le classifieur vidéo propose 7 classes alors que le classifieur audio propose deux classes seulement) ? Faut-il faire une classification à partir de chaque type d'information puis fusionner les décisions obtenues ou alors faut-il faire une classification qui tienne compte de toutes les informations ?

6 Interprétation des gestes faciaux

Beaucoup d'informations liées à la communication sont contenues dans les gestes faciaux. Par exemple, il est courant de secouer verticalement et périodiquement la tête en signe d'approbation et de secouer horizontalement et périodiquement la tête en signe de négation. Autre exemple, le froncement des sourcils est courant quand quelqu'un est dubitatif. Concernant l'analyse des expressions faciales, une expression donnée s'accompagne elle aussi bien souvent de gestes typiques de la tête. Par exemple, lorsque quelqu'un se sent peu à l'aise ou honteux, il a tendance à baisser la tête.

Ces gestes faciaux font partie intégrante de notre système de communication comme en témoigne certaines expressions telles que « être bouche-bée » par exemple. Si on s'intéresse à l'analyse de la langue des signes, langage le plus répandu pour la communication entre personnes sourdes, on s'aperçoit que l'intégration de l'information à transmettre au travers de gestes faciaux est poussée à l'extrême. C'est un moyen de compenser et de suppléer le canal audio qui est inexploitable voire inexistant. Prenons l'exemple de la Figure 13 issu de [Ong05]. Sur cet exemple de la langue des signes américaine, la main code deux mots : GIRL et HERE. Ce sont alors les gestes faciaux qui permettent de créer trois phrases différentes avec ces deux mots : à savoir une forme affirmative « *the girl is here* », une forme interrogative « *is the girl here ?* » et une forme négative « *the girl is not here* ». Même sans

aucune connaissance particulière de la langue des signes américaine, l'examen des trois couples d'images nous permet de deviner chacune des trois formes. En effet, dans le premier cas, les gestes de la main s'accompagnent d'un hochement de tête d'approbation ; dans le second cas, les gestes de la main s'accompagnent d'un écarquillement des sourcils ainsi que d'un mouvement global vers l'avant de la tête et des épaules ; dans le dernier cas, les gestes de la main s'accompagnent d'un hochement de tête de négation.

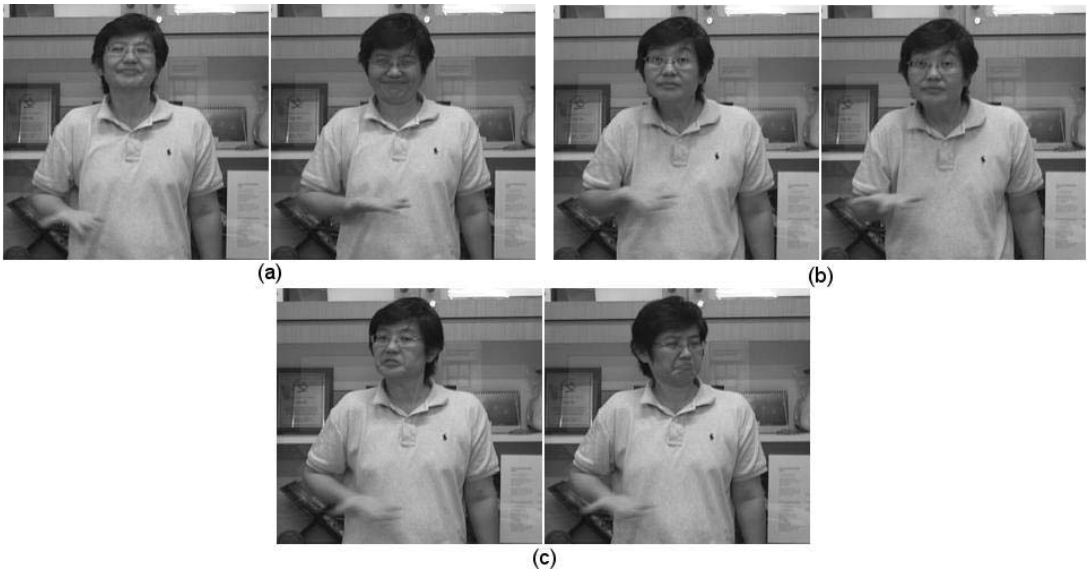

Figure 13 : a) forme affirmative ; b) forme interrogative ; c) forme négative [Ong05]

L'objet des travaux développés dans cette partie consiste à développer des méthodes d'analyse et d'interprétation de gestes faciaux. On s'intéresse plus particulièrement aux mouvements rigides de la tête (hochements…) et aux mouvements associés aux traits du visage (clignements volontaires ou non des yeux, bâillement, écarquillement des yeux, froncement de sourcils …). La caractéristique principale de l'approche adoptée est qu'elle s'appuie sur des processus issus de la modélisation du fonctionnement du système visuel humain.
Les travaux décrits dans cette partie sont développés dans le cadre de la thèse d'Alexandre Benoit dont la soutenance est prévue pour l'automne 2006.

6.1 Analyse de mouvement inspiré du fonctionnement du système visuel humain

Depuis plusieurs années, sous la direction de Jeanny Hérault, le groupe *perception* du laboratoire LIS s'intéresse à la modélisation du fonctionnement du système visuel humain [Beaudot94, Torralba99, Alleyson99]. En particulier, il a été mis en évidence qu'un pré-traitement spécifique est réalisé au niveau de la rétine en vue de l'analyse des stimuli en mouvement. Nous avons utilisé les résultats de ces travaux comme point de départ pour le développement d'un algorithme nouveau d'analyse et d'interprétation des mouvements rigides et non rigides du visage.
Dans une première étape, chaque image de la vidéo traitée est filtrée au moyen d'un filtre modélisant les traitements ayant lieu au niveau de la rétine humaine. Dans un second temps, le spectre des images filtrées est calculé dans le domaine log-polaire afin de modéliser les traitements ayant lieu au niveau du cortex visuel primaire. L'analyse du spectre obtenu permet de remonter à des informations à propos des mouvement présents dans la scène analysée.

6.1.1 Filtrage rétinien

Le principe de la méthode d'analyse des mouvements de la tête a été décrit dans **[CI_Benoit05a, CI_Benoit05b]**. Au niveau de la rétine s'effectue un pré-traitement des images qui se déroule en deux phases [Beaudot94] :

- au niveau de la couche plexiforme externe (Outer Plexiform Layer ou OPL), tous les traitements peuvent être modélisés par un filtre spatio-temporel non séparable (cf. Figure 14). A basse fréquence temporelle, ce filtre a un effet passe-bande spatial ce qui a pour effet l'accentuation des contours présents dans l'image. A basse fréquence spatiale, ce filtre a un comportement de type passe-bande à large bande passante ce qui a pour effet une atténuation des variations locales d'illumination. A haute fréquence temporelle, le filtre se comporte comme un filtre passe-bas ce qui permet une atténuation du bruit spatio-temporel.
- au niveau de la couche plexiforme interne (Inner Plexiform Layer ou IPL), il existe un processus dédié à la détection des stimuli en mouvement. Ce processus peut être modélisé par un opérateur simple de dérivation temporelle [Ritcher82]. Ceci a pour effet, une accentuation des contours en mouvement au détriment des contours statiques. La réponse en sortie de l'IPL est d'autant plus importante au niveau des contours perpendiculaires au mouvement.

La Figure 15 présente un exemple de résultat obtenu successivement à la sortie de l'OPL puis de l'IPL. L'image proposée est une image extraite d'une séquence où la tête effectue un mouvement de tangage (tilt motion). A la sortie de l'OPL, l'ensemble des contours a été accentué. En revanche, à la sortie de l'IPL, seuls apparaissent fortement les contours dont l'orientation est perpendiculaire au mouvement. Cette méthode est particulièrement intéressante pour l'analyse des mouvements de la tête car les contours y sont majoritairement verticaux ou horizontaux.

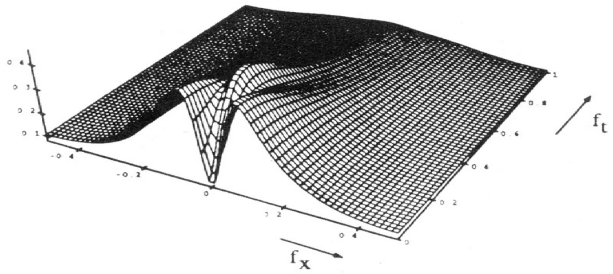

Figure 14 : réponse en fréquence du filtre modélisant le comportement de l'OPL [Beaudot94]

Figure 15 : de gauche à droite : image d'une séquence de tilt ; sortie de l'OPL ; sortie de l'IPL

Un autre intérêt de cette approche est le réhaussement des contours en mouvement se fait en temps réel. En effet, le filtre OPL est modélisé par une succession de 4 filtrages directionnels récursifs causaux et anti-causaux effectuant chacun 2 opérations par pixel [Beaudot94].

6.1.2 Cortex visuel primaire : FFT et transformée log-polaire.

Après l'étape de filtrage rétinien, le flux vidéo contient essentiellement les contours en mouvement. Le spectre de chaque image est calculé et une transformation log polaire [Oliva99] est réalisée. Lors de la transformation log-polaire, on calcule la réponse du spectre à un ensemble de filtres passe-bandes orientés (couramment des filtres de Gabor). La transformation log polaire permet d'identifier un zoom comme une translation des énergies spectrales le long de l'axe des fréquences. De même, une rotation dans le plan de la caméra (rotation 2D simple) est équivalente à une translation des énergies spectrales sur l'axe des orientations. Ceci est illustré sur la Figure 16.

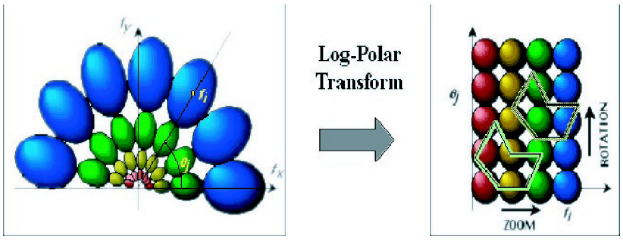

Figure 16 : effet de la transformée log-polaire

6.2 Estimation des mouvements de la tête

A partir de l'analyse du spectre log-polaire, il est possible d'en déduire le type de mouvement (translation ou rotation) ainsi que sa direction. Dans les travaux proposés ici, nous ne nous sommes pas focalisés sur le problème de l'estimation précise du module de la vitesse. En effet, notre objectif étant l'interprétation haut niveau (analogue aux interprétations humaines) des mouvements de la tête, l'amplitude n'apparaît pas comme une donnée fondamentale.

6.2.1 Interprétation du spectre log-polaire

Direction du mouvement

Le spectre log polaire présente les plus fortes énergies au niveau des fréquences associées aux contours perpendiculaires au mouvement. Par exemple, la Figure 17 montre le spectre log polaire d'un objet en translation verticale à texture uniformément orientée. L'énergie est concentrée autour de l'axe vertical (orientation 90°).

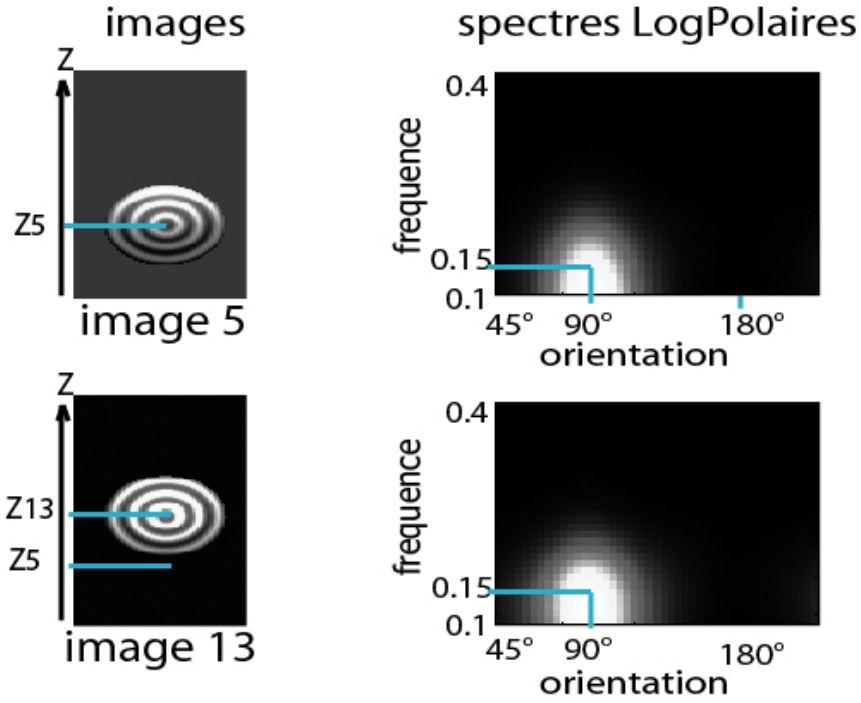

Figure 17 : évolution du spectre d'un objet en translation

Pour estimer la direction du mouvement, nous cumulons les énergies des réponses des filtres pour chaque orientation. Ceci nous amène à une courbe d'énergie cumulée par orientation dont la Figure 18 a-b-c-d montre plusieurs exemples pour différents mouvements de tête. Sur chacune de ces courbes, l'abscisse du maximum d'amplitude correspond à l'orientation des contours les plus énergétiques, à savoir ceux perpendiculaires au mouvement global. Les Figure 18 a-b-c montrent que lors d'un mouvement simple, il n'apparaît qu'un seul maximum sur la courbe d'énergie cumulée par orientation. De l'abscisse de ce maximum, on déduit aisément l'orientation du déplacement. Dans le cas d'une rotation composée comme celle de la Figure 18-d, on constate que la courbe d'énergie possède 2 maximums, d'où 2 mouvements élémentaires.

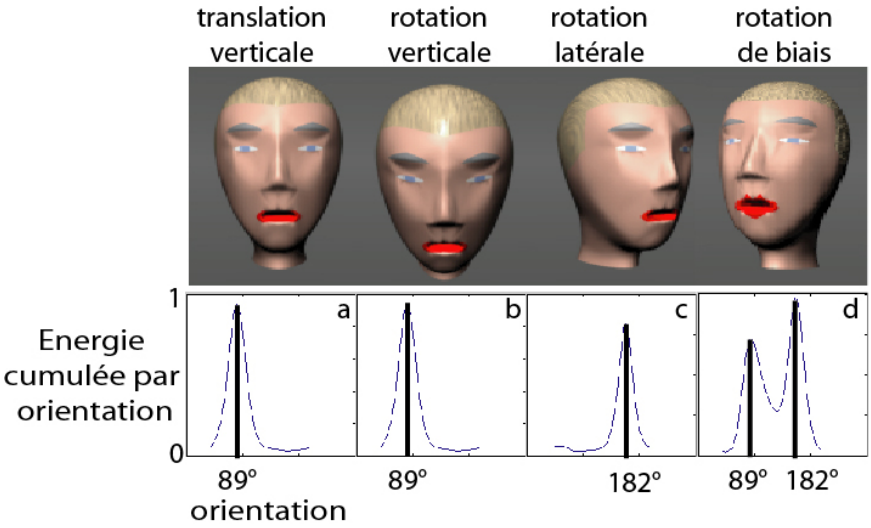

Figure 18 : mouvements simples et composés. En haut, différents mouvements pour un visage de synthèse ; en bas, courbe d'énergie cumulée par orientation.

Type de mouvement

Le type de mouvement (rotation ou translation) est lié à la position fixe ou mobile au cours du temps du maximum d'énergie du spectre log polaire, et donc à l'évolution temporelle de la position du maximum de la courbe d'énergie cumulée par orientation. Zoom et rotation 2D (*roll*) induisent des translations globales du spectre, respectivement selon l'axe des fréquences et des orientations. Les rotations 3D (*pan* et *tilt*) induisent des translations locales du spectre. Enfin, les translations dans le plan image sont caractérisées par un spectre invariant temporellement.

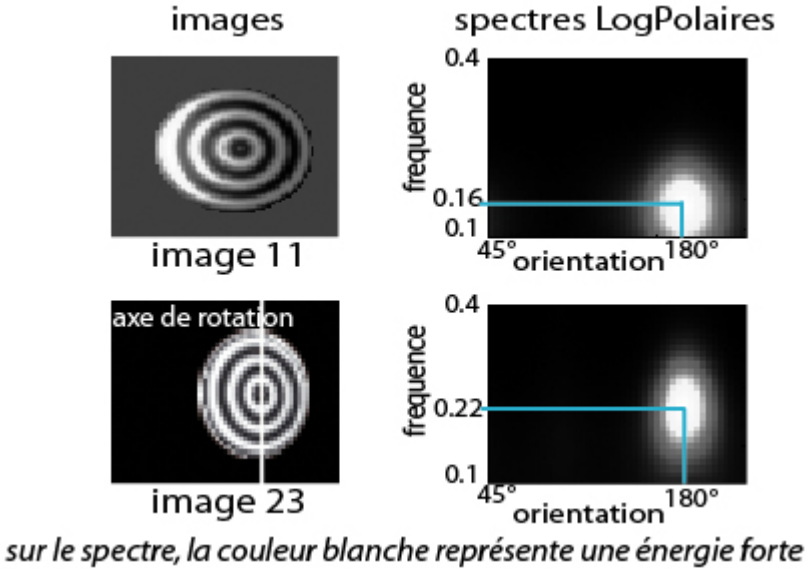

Figure 19 : évolution de l'énergie du spectre log-polaire d'un objet en rotation

On constate par exemple sur le spectre log-polaire de la Figure 17 qu'il n'y a pas de déplacement du maximum d'énergie au cours d'une translation. En effet, lors d'une translation, les contours perpendiculaires au mouvement ne sont pas modifiés. La Figure 19 quant à elle montre l'évolution du spectre log-polaire sur un objet de texture orientée uniformément sur toutes les orientations et subissant une rotation selon un axe vertical. Les contours extraits à l'issu du pré filtrage sont principalement ceux alignés avec l'axe de rotation. Or, du fait de la rotation 3D, ces mêmes contours se compressent entre eux sur le plan image ce qui entraîne une augmentation de leur fréquence associée. Sur cet exemple, de l'image 11 à l'image 23, l'objet a subi une rotation de 25° ce qui entraîne ici une compression des contours verticaux. Leur fréquence normalisée passe de f_{11}=0.16 à f_{23}=0.22. On peut donc conclure qu'une rotation 3D (de type *pan* ou *tilt*) se caractérise d'une part par une énergie maximum pour les contours alignés avec l'axe de rotation, et d'autre part par des translations d'énergie sur l'axe fréquentiel.

Amplitude du mouvement

Bien que l'estimation précise de l'amplitude du mouvement ne soit pas notre préoccupation première lorsqu'il s'agit d'appréhender les « gestes » faciaux, nous avons fait une étude sur l'évolution de l'énergie du spectre log polaire en fonction de l'amplitude du mouvement (cf. Figure 20). L'énergie du spectre log-polaire apparaît donc proportionnelle à la vitesse sauf à faible vitesse. En effet, en absence de mouvement, l'énergie du spectre log-polaire est nulle puisqu'il n'y a plus aucun contour en mouvement. Cette propriété permet de détecter automatiquement les arrêts de mouvement ainsi que les changements de sens de mouvement.

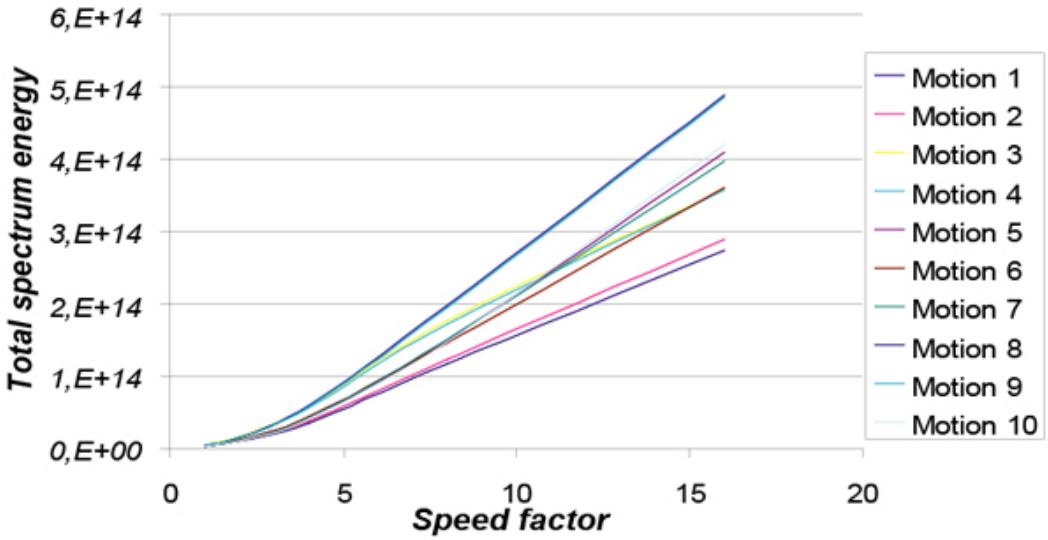

Figure 20 : évolution de l'énergie du spectre log-polaire en fonction de l'amplitude de la vitesse

6.3 Interprétation haut niveau

Le détecteur de mouvement décrit précédemment permet d'obtenir des informations sur la nature des mouvements de la tête. A partir de là, nous pouvons envisager une analyse de ces mouvements à un plus haut niveau afin de leur associer une interprétation : signification de certains hochements de tête, tête lourde, désignation d'une région d'intérêt…

6.3.1 Hochements de tête : gestes de communication non verbale

Dans ce paragraphe, nous nous intéressons à la détection automatique des hochements de tête associés à l'approbation et à la négation. Ces hochements sont des informations de communication en face à face. Ceci a fait l'objet des publications **[CN_Benoit05, CI_Benoit05c]**. L'analyse plus précise des hochements considérés montre que :

- un hochement d'approbation se traduit par un mouvement périodique de rotation de type tangage (ou *tilt*) de la tête. Plus précisément, il faut détecter une rotation de type *tilt* (direction 90°) avec des changements périodiques de direction.
- un hochement de négation se traduit par un mouvement périodique de rotation de type lacet (ou *pan*) de la tête. Plus précisément, il faut détecter une rotation de type *pan* (direction 180°) avec des changements périodiques de direction.

Afin de détecter chacun de ces deux types de hochement de tête, l'évolution temporelle du maximum de l'énergie cumulée par orientation ainsi que celle de l'énergie totale du spectre à la sortie de l'IPL sont analysées. La Figure 21 montre des exemples de telles courbes pour une séquence dans laquelle le sujet effectue des hochements d'approbation de l'image 370 à 400 et des hochements de négation entre les images 400 et 462. La courbe gauche de la Figure 21 montre l'évolution temporelle de l'orientation associée au maximum de la courbe d'énergie cumulée par orientation. On constate que les orientations associées aux maxima détectés sont successivement 90° entre les images 371 et 398 (correspond au mouvement de *tilt* des contours horizontaux) puis 180° entre les images 400 et 463 (correspond au mouvement de *pan* des contours verticaux). La courbe droite de la Figure 21 montre l'évolution temporelle de l'énergie totale du spectre pour cette même séquence vidéo. On constate que cette énergie se présente sous la forme d'un signal périodique avec des maxima et minima à intervalles réguliers. Sachant que les minima sont associés aux mouvements lents voire aux arrêts de la tête, on enregistre donc sur cette séquence une succession de mouvements et d'arrêts.

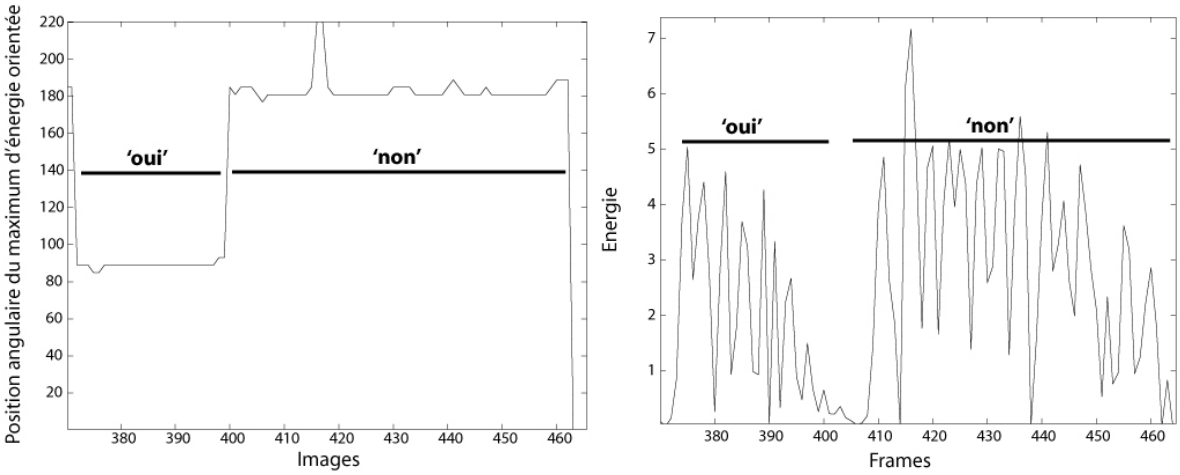

Figure 21 : à gauche, évolution temporelle de la position angulaire du maximum de la courbe d'énergie cumulée par orientation ; à droite, évolution temporelle de l'énergie totale du spectre.

L'algorithme proposé pour l'identification d'une approbation ou d'une négation est le suivant :
Soient α_m et α_d respectivement la moyenne et l'écart type de l'orientation des *n* derniers maximums rencontrés (couramment, *n*=3) et τ_m et τ_d la moyenne et l'écart type de la période d'apparition de ces maximums. Soit ε le rapport entre la moyenne du bruit spatio temporel et la moyenne des *n* maximums d'énergie.

si ε< 0.5 (le maximum courant d'énergie est différent du bruit)
si τ_m <1/f_{min}=0.83s et τ_d <0.2 (période stable, >0.83s, au dessous mouvements trop lents)
si α_m=90° et α_d < 15° (mouvement d'orientation proche de la verticale)
alors «Approbation»
sinon si α_m=180° et α_d< 15° (mouvement d'orientation proche de l'horizontale)
alors «Négation»
sinon «Attitude indéterminée»
sinon «Attitude indéterminée»
sinon «Attitude indéterminée»

L'attitude *indéterminée* représente toute attitude différente d'un hochement d'approbation ou de négation. Le ratio ε permet de limiter les fausses détections dues au bruit.
Le système détecte les approbations et négations avec un taux réussite de 100% en condition standard d'éclairement (éclairage de bureau) pour un visage occupant de 20% à 100% de la taille des images. La résistance au bruit a été évaluée, dans des conditions d'éclairement faible ou de bruit (test avec bruit gaussien de variance 0.05). L'algorithme détecte les approbations et négations avec 90% de réussite. Par ailleurs, la sensibilité aux occultations partielles du visage a été testée : le taux de réussite est de 80% si le visage est caché à 50%. Ceci est dû au fait que l'analyse fréquentielle nécessite seulement un échantillon de contours du visage regroupant les principales orientations.

6.3.2 Analyse de vigilance : états des yeux, fréquence de clignement et détection de bâillement.

Notons que la seule contrainte vis à vis du détecteur proposé est l'existence de contours perpendiculaires au mouvement. De ce fait, nous utilisons la même méthode pour l'analyse des mouvements de certains traits du visage tels que les yeux et la bouche. En effet, un clignement est associé à un mouvement vertical rapide de la paupière et un bâillement à un mouvement vertical rapide de la bouche. L'état ouvert ou fermé de chacun de ces deux traits

est à relier à la valeur de l'énergie à la sortie de l'OPL qui est plus élevée pour un trait ouvert que pour un trait fermé. Ce travail a fait l'objet de la publication **[CI_Benoit05d]**. Par ailleurs, ceci a également été intégré dans le projet de « simulateur de conduite avec détection des états d'hypovigilance » que nous avons développé dans le cadre du Workshop eNterface **[project4]** (cf. 9.4.3).

Détection des clignements et des bâillements.

On suppose ici qu'il est possible d'extraire une boite englobante autour de chaque œil et de la bouche, ceci étant fait par exemple grâce aux travaux décrits au paragraphe 5.1. Sur chacune de ces boîtes, on place le détecteur de mouvement décrit précédemment. L'énergie à la sortie de l'IPL étant proportionnelle au mouvement des contours présents, il apparaît un fort pic d'énergie à chaque clignement ou à chaque bâillement. La Figure 22-gauche présente l'évolution de l'énergie à la sortie de l'IPL pour une séquence vidéo constituée de plusieurs clignements successifs. On constate sur cette courbe que tous les clignements conduisent certes à un pic d'énergie mais ce pic peut être d'amplitude différente selon les clignements. Afin de s'affranchir de cette variabilité et d'être capable de détecter de manière fiable tous les clignements présents, un indicateur de fiabilité du mouvement présent, tenant compte entre autre du niveau de bruit présent dans l'image et indépendant du niveau absolu de l'énergie à la sortie de l'IPL, est calculé. La Figure 22-droite montre l'évolution temporelle de cet indicateur noté α_1. Pour chaque clignement, cet indicateur normalisé est à 1, alors que lorsque l'œil reste ouvert et qu'il n'y a pas de mouvement, cet indicateur est proche de 0. Ceci permet de détecter de manière fiable tous les clignements (pour peu que la cadence de traitement soit suffisante). On procède de la même manière pour la détection des bâillements.

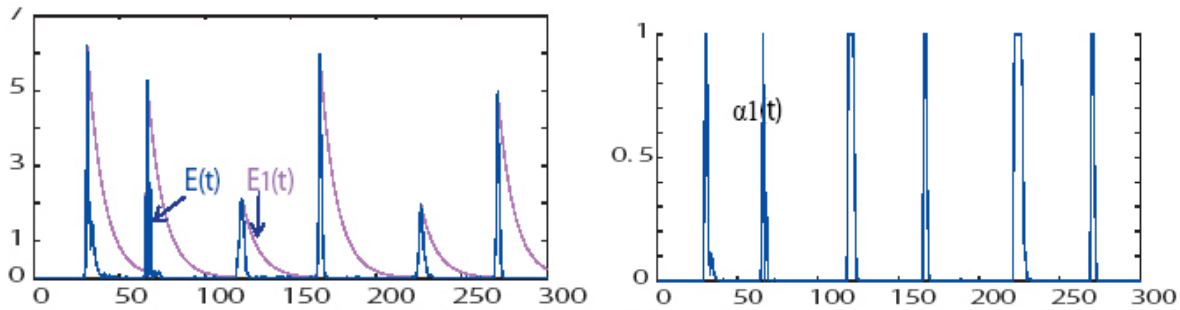

Figure 22 : à gauche, évolution temporelle de l'énergie E(t) à la sortie de l'IPL ; à droite, évolution de l'indicateur d'événement.

Détection de l'état ouvert ou fermé de la bouche et des yeux.

La Figure 23 présente pour chaque trait et chaque configuration possible le niveau d'énergie associé à la sortie de l'OPL. Ceci confirme le fait que cette énergie est plutôt élevée pour les traits ouverts (présence de beaucoup de contours horizontaux) et est de valeur plutôt faible pour les traits fermés (disparition de certains contours). Cependant la valeur absolue du niveau d'énergie à la sortie de l'OPL pour chacun des traits dépend de la personne considérée. D'où l'idée de coupler ce détecteur d'état avec le détecteur de mouvement afin d'apprendre et éventuellement de remettre à jour le niveau d'énergie associé à chaque état et à chaque trait ; cette remise à jour se faisant à chaque fois qu'un événement (clignement ou bâillement) a été détecté.

	OPL Output	Total Energy	Binary state
Eye open		4.7	High energy level
Eye closed		2.8	Low energy level
Mouth open		1.9	High energy level
Mouth closed		1.1	Low energy level

Figure 23 : Etats possibles pour les yeux et la bouche et énergies associées

6.4 Ce qu'il reste à faire

Nous avons développé un algorithme performant pour l'analyse des mouvements rigides et non rigides du visage. Des systèmes simples d'interprétation et de reconnaissance des gestes faciaux associés à ces mouvements ont été proposés. Par la suite, nous envisageons le développement de systèmes plus complexes de reconnaissance multimodale comme par exemple la fusion des informations obtenues lors de la phase d'analyse des expressions faciales avec celles obtenues lors de l'analyse des mouvements du visage. Nous sommes convaincus que l'apport de ces nouvelles informations permettra d'augmenter les performances de l'algorithme de reconnaissance des expressions faciales, en particulier en levant la contrainte de tête fixe. Néanmoins, une telle fusion ne sera pas immédiate. De nombreuses questions se posent et en particulier, faut-il faire une fusion de données au niveau des informations bas niveau extraites (déformations des contours des traits du visage + type de mouvement facial) ou au niveau des décisions qui ont déjà été prises par chacun des systèmes pris séparément (groupe d'expressions les plus probables + bâillement ou fermeture des yeux…) ?

Un autre point à creuser consiste en l'étude des performances du détecteur de mouvement basé sur la rétine pour l'analyse de mouvements quelconques (autres que des mouvements liés au visage). En effet, la seule hypothèse forte imposée par la méthode proposée est l'existence dans la scène analysée de contours perpendiculaires au mouvement présent. Du fait de la structure géométrique du visage (présence des yeux, de la bouche, du nez…), cette hypothèse est valide pour l'analyse des mouvements du visage. Mais il est fort probable que cet algorithme est beaucoup plus générique et qu'il pourrait permettre l'analyse de mouvements autres que ceux associés à la tête.

7 Reconnaissance de postures

L'analyse du comportement humain dans les séquences vidéo, étroitement liée à l'analyse et l'interprétation de ses mouvements, est un thème de recherche en pleine expansion. Les articles [Aggarwal03, WangJ03, WangL03] proposent des synthèses sur tous les travaux récents effectués dans ce domaine. Cet engouement est justifié par la multiplicité des applications possibles telles par exemple la vidéo surveillance (voir le projet *VSAM : Video Surveillance And Monitoring* [Collins00], projet « W^4 *: Who ? When ? Where ? What ?* » de détection, suivi et surveillance d'activités humaines [Haritoaglu98]), le développement de systèmes virtuels interactifs à réalité mixte [Wreng97], l'amélioration des interfaces homme machine…

Nous nous sommes intéressés au problème de reconnaissance de postures avec deux motivations essentielles :

- en parallèle de ce que nous faisons sur l'analyse des mouvements de tête et des mouvements des traits du visage, nous voulons compléter nos travaux sur l'analyse des mouvements humains en s'intéressant au mouvement plus global d'une personne. Cela permet alors de disposer d'une analyse du comportement d'une personne à plusieurs échelles : d'un point de vue global en s'intéressant aux mouvements de l'ensemble de la personne et d'un point de vue plus local en se focalisant sur les mouvements de la tête et du visage.
- deux applications particulières nous intéressent dans ce cadre : le développement de systèmes de réalité mixte (cf. le projet européen Artlive §9.1), la surveillance de personnes âgées ainsi que le développement de systèmes ambiants intelligents.

Dans les travaux présentés dans cette section, l'objectif est de reconnaître automatiquement qu'une personne (sous surveillance vidéo) se trouve dans l'une des quatre postures suivantes : *assise*, *debout*, *couchée*, *accroupie*. Comme pour l'interprétation des mouvements de tête ou pour la reconnaissance d'expressions faciales, la reconnaissance de postures requiert deux phases essentielles : d'une part, une phase d'extraction automatique d'informations pertinentes dans la séquence vidéo (cf. section 7.1) et d'autre part, une phase de fusion de ces informations pour en déduire la posture adaptée (cf. section 7.2).
Ces travaux ont été développés dans le cadre des DEA de N. Mottin, Y.Lauriou, V. Girondel , J. Romeuf et N. Cacciaguera puis très largement complétés dans le cadre de la thèse de V. Girondel dont la soutenance est prévue pour l'automne 2005.

7.1 Extraction d'indices bas niveau : détection et suivi de personnes, identification de la tête et des mains

7.1.1 Détection de personnes par approche markovienne.

Pour la segmentation des personnes en mouvement, nous avons utilisé et amélioré l'algorithme proposé dans le cadre de ma thèse**.** Ceci a conduit à la publication **[CI_Caplier01a]**. La détection de personnes en mouvement devant une caméra fixe est envisagée dans un cadre markovien comme la recherche du champ d'étiquettes binaires (chaque pixel est soit fixe soit mobile) le plus probable étant donné un ensemble d'observations. Les deux observations considérées sont d'une part la différence inter-image qui met en évidence les variations temporelles de la fonction de luminance et d'autre part, l'utilisation d'une image de référence (image du fond fixe de la scène analysée) construite et mise à jour dynamiquement. Le théorème de Bayes montre que la recherche de ce champ d'étiquettes le plus probable est équivalente à la minimisation d'une fonction d'énergie faisant intervenir :

- un terme d'énergie a priori qui confère alors des propriétés a priori (homogénéités spatiale et temporelle) sur le champ recherché ;
- un terme d'énergie d'adéquation aux observations qui permet de s'assurer que la solution trouvée est en accord avec les données images.

La minimisation de cette énergie totale étant itérative, nous avons remplacé certaines de ces itérations par des opérateurs morphologiques ce qui nous a permis de réduire le temps de calcul pour atteindre la solution.
La Figure 24 donne des exemples de segmentation. La séquence présentée à gauche concerne une scène d'intérieur avec un unique personnage en mouvement. Dans ce cas, l'obtention de l'image de référence est très simple, on se contente de filmer la scène avant l'arrivée du personnage. Pour la séquence du milieu qui représente une scène de rue, il a fallu construire pixel à pixel l'image de référence et la mettre à jour afin de tenir compte des éventuelles

variations d'éclairement. Enfin, la dernière séquence présentée montre que l'algorithme proposé n'est pas dédié uniquement à la détection de personnes. Il peut fonctionner pour détecter tout type d'objets en mouvement.

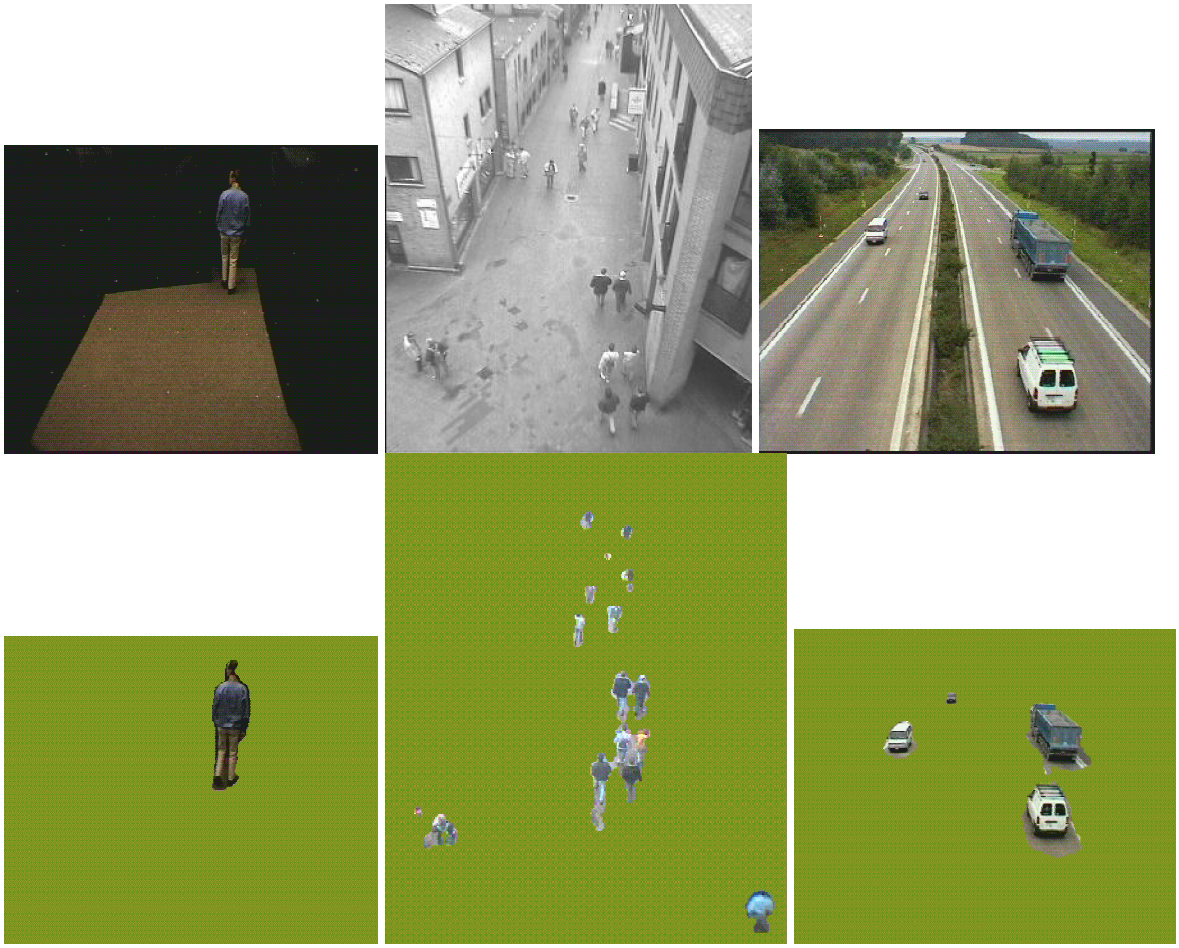

Figure 24 : exemples de résultats de détection de mouvement. En haut, une image extraite de chaque séquence ; en bas, masques des objets détectés en mouvement.

7.1.2 Suivi de personnes par filtrage de Kalman à données incomplètes

Une fois détectés les personnages en mouvement dans une scène, nous nous sommes intéressés à leur suivi au cours du temps. Il s'agit d'établir un lien temporel entre les masques successifs obtenus. La difficulté de cette tâche réside essentiellement en la gestion du suivi en cas d'occultation partielle ou totale d'une personne.

Principe de base du suivi

Pour des raisons de complexité de calcul, le suivi a été défini non pas sur le masque des personnes mobiles mais sur la boite rectangulaire englobant ces masques. Nous avons développé un algorithme de suivi qui procède par analyse de l'intersection de ces boites englobantes au cours du temps. Le principe de base est de mettre en correspondance temporelle les boites englobantes ayant le plus grand recouvrement. Néanmoins, pour gérer des cas d'apparition ou de disparition d'objets, de fusion ou de séparation dans un groupe, on établit la liste des prédécesseurs possibles (le prédécesseur le plus probable étant celui qui a le recouvrement le plus grand) ainsi que la liste des successeurs possibles et on confronte ces deux listes jusqu'à trouver une mise en correspondance qui soit compatible dans les deux sens (avant et arrière) ce qui permet d'établir le lien temporel. Il existe toujours une solution sauf

dans le cas de l'apparition d'une nouvelle personne. L'absence de lien temporel est donc interpréter comme l'apparition d'une nouvelle personne.
Une fois le lien temporel établi, il est facile de tracer la trajectoire du centre de gravité de chaque personne. La Figure 25 présente un exemple de suivi et de tracé de trajectoire pour trois personnes différentes.

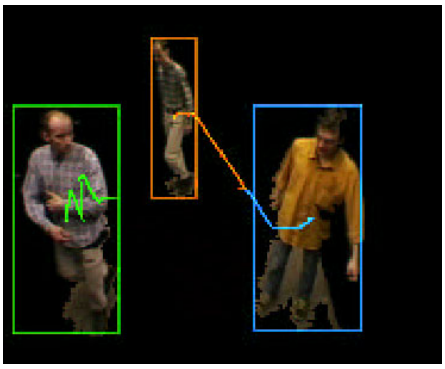

Figure 25 : suivi de personnes et tracé de trajectoires

Gestion des occultations : filtrage de Kalman

La méthode présentée précédemment ne permet pas la gestion des occultations. Pour ce faire, nous avons utilisé un filtrage de Kalman qui permet de faire une prédiction non seulement de la boîte englobante mais aussi de la position et de la vitesse de la tête d'une personne dans l'image suivante (cf. **[CI_Girondel04]**). Dans cette partie, nous supposerons que nous sommes capable d'extraire une boite englobante autour de la tête de chaque personne en mouvement (voir §7.1.3 pour la description de la méthode). Nous supposons aussi que nous sommes capable d'estimer la vitesse de la tête de chaque personne que nous assimilons à la vitesse de la personne. Cette estimation de vitesse est faite par mise en correspondance ou *block matching*.
Le modèle choisi est un filtre de Kalman à modèle d'évolution à vitesse constante. A priori, le filtrage travaille sur 10 mesures (4 coordonnées pour la boite englobant la personne, 4 coordonnées pour la boite englobant sa tête et 2 valeurs pour la vitesse de la tête). En cas d'occultation partielle, certaines de ces mesures peuvent être non disponibles. Le filtre de Kalman travaille en mode prédiction et remise à jour sur les données disponibles et en mode prédiction uniquement sur les données non disponibles.
La Figure 26 présente un exemple de résultats de suivi de personnes en cas d'occultation. Sur les résultats proposés, les boites en traits pleins correspondent aux boites détectées et les boites en pointillées correspondent aux boites prédites par Kalman. Avant le regroupement (image 200) et après la séparation des deux personnes (images 228 et 231), toutes les mesures nécessaires au filtrage de Kalman sont disponibles (boite englobant la personne + boite englobant la tête + vitesse de la tête). Pour les images 212, 213 et 219, le filtre de Kalman est en mode prédictif pour l'une des deux têtes qui est occultée.

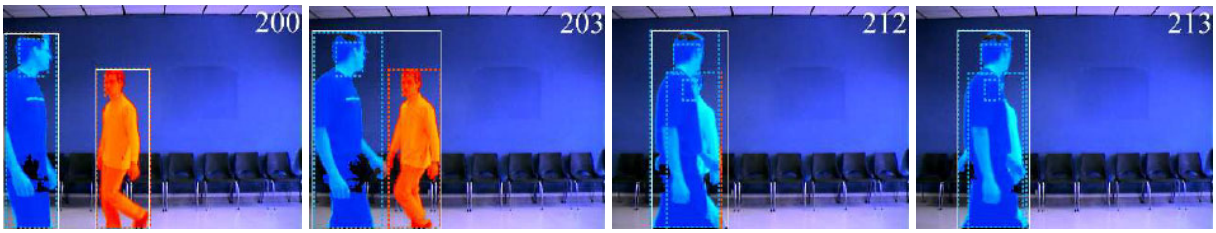

Figure 26 : exemple de suivi de personne en cas d'occultation

7.1.3 Détection des pixels de peau : identification et suivi du visage et des mains

L'objectif est de pouvoir suivre tout au long d'une séquence le visage et les mains d'une personne. Le suivi du visage permet de résoudre le cas du suivi d'une personne en cas d'occultation partielle de son corps. De plus, la distance entre le visage et les pieds d'une personne est utilisée dans l'algorithme de reconnaissance de postures décrit au §7.2. La connaissance de la position des mains est une donnée intéressante pour les applications interactives de réalité mixte. Ces travaux ont été décrits dans **[CI_Girondel02].**

Détection des pixels de peau

De nombreux travaux ont mis en évidence la spécificité de la chrominance des pixels de peau [Terrillon00]. En se plaçant dans un espace à luminance et chrominance séparées, on montre que les différences entre toutes les couleurs de peau (au sens commun du terme couleur) ne sont contenues que dans la luminance. Dans la littérature, il n'existe pas d'espace chrominance qui fait l'unanimité pour la détection des pixels de peau, plusieurs sont couramment utilisés tels que l'espace RGB normalisé, l'espace HSV, l'espace YIQ… Nous avons donc fait notre propre analyse sur une double base d'échantillons de peau : la base de Von Luschan (voir Figure 27-gauche) et une base d'échantillons de pixels de peau que nous avons acquis au laboratoire. Sur ces données, c'est l'espace CbCr qui est apparu le plus discriminant. La Figure 27-droite montre la répartition des pixels de nos échantillons de peau dans cet espace. Nous extrayons alors les pixels peau dans une image par un double seuillage : $C_b \in [90,130]$ *et* $C_r \in [130,180]$. Dans le choix de l'espace couleur, nous avons également tenu compte de la contrainte de complexité de calcul en plus du pouvoir discriminant. En effet, en prenant l'espace CbCr, nous pouvons utiliser directement les images issues de la caméra vidéo (aucune conversion de format n'est nécessaire) et l'extraction des pixels de peau est simple puisqu'elle se fait par double seuillage.

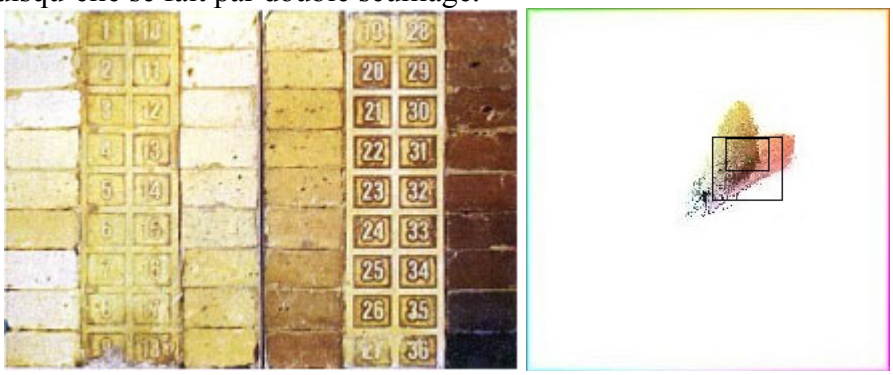

Figure 27 : à gauche, base de peau de Von Luschan ; à droite, répartition des pixels de peau de la base d'apprentissage dans l'espace CbCr

Cependant, il apparaît que les seuils doivent être affinés afin de s'adapter au système d'acquisition et à une personne particulière. Les gammes de valeurs fournies précédemment

servent donc de valeurs initiales et elles sont mises à jour au cours de la séquence par un processus de translation et d'homothétie appliqué au rectangle initial (le grand rectangle noir initial de la Figure 27 se transforme au cours de la séquence pour aboutir au petit rectangle noir de cette même figure après adaptation à une caméra et une personne).
La Figure 28 présente un exemple de résultats.

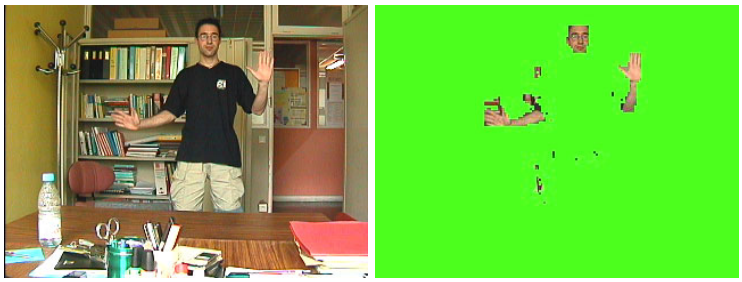

Figure 28 : à gauche, une image ; à droite, pixels de peau extraits

Identification et suivi du visage et des mains

Après un étiquetage en composantes connexes des pixels détectés comme pixels de peau, les trois tâches correspondant aux deux mains et au visage sont détectées et suivies au cours du temps par la prise en compte d'un ensemble de critères spatiaux et temporels triés dans un ensemble de listes. L'identification de chaque composante a lieu en tenant compte d'un ensemble de connaissances a priori. Par exemple, la tête est supposée correspondre en général à la tâche la plus grosse, la plus haute, la plus proche la position détectée pour la tête dans l'image précédente. La Figure 29 présente des exemples de localisation du rectangle englobant chaque main ainsi que le visage. Un code de couleur a été attribué à chaque composante. La détection et la localisation s'appuyant sur la localisation de tâches connexes de pixels de couleur peau, les avant bras sont aussi détectés dans le cas où la personne filmée est bras nus.

Figure 29 : localisation des mains (cadre bleu pour la main droite et cadre rouge pour la main gauche) et de la tête (cadre vert)

7.2 Système de reconnaissance de postures statiques : assis, debout, accroupi, couché

La finalité de ces travaux étant la reconnaissance de postures, nous avons intégré les informations bas niveau extraites sur les vidéos dans un processus de fusion de données utilisant le théorie de l'évidence afin de reconnaître les quatre postures statiques suivantes :

- hypothèse H_1 : *debout* ;
- hypothèse H_2 : *assis* ;
- hypothèse H_3 : *accroupi* ;
- hypothèse H_4 : *couché*.

Ceci a conduit aux publications **[CI_Girondel05a, R_Girondel05, CI_Girondel05b, CI_Girondel05c]**.

Mesures ou données considérées

Afin de reconnaître chacune des 4 postures considérées, trois distances normalisées sont définies (cf. Figure 30), elles résultent de l'exploitation des indices bas-niveau extraits grâce aux méthodes décrites à la section 7.1:

- $r_1 = D_1 / D_1^{ref}$: représente la distance normalisée du centre du visage jusqu'au bas de la boite englobante rectangulaire entourant la personne détectée;
- $r_2 = D_2 / D_2^{ref}$ représente la hauteur normalisée de la boite englobante rectangulaire;
- $r_3 = D_3 / D_3^{ref}$ représente la distance normalisée du centre du visage au centre de gravité de la boite englobante par axes principaux d'inertie ;
- $r_4 = D_4 / D_4^{ref}$ représente la longueur normalisée du demi grand axe de la boite englobante par axes principaux d'inertie.

La normalisation est effectuée en considérant une posture de référence où la personne est debout avec les bras écartés à l'horizontal (cf. Figure 30-droite).

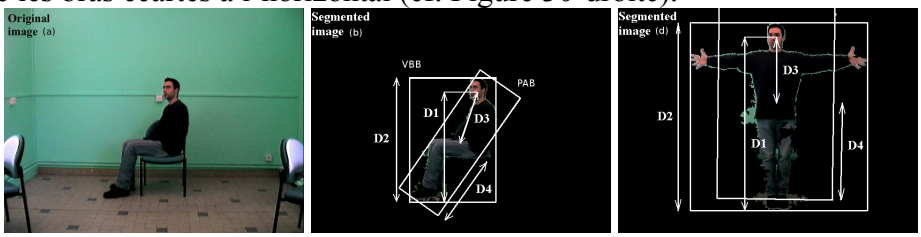

Figure 30 : de gauche à droite : image initiale, personnage segmenté avec boites englobantes et distances considérées, posture de référence avec définition des distances de référence

Modélisation et fusion de données

L'analyse de l'évolution des mesures r_i pour les diverses postures considérées a conduit à la définition de deux modèles (cf. Figure 31) :

- un modèle pour r_1 *et* r_2 : l'idée de ce modèle est que plus le visage est bas, plus il y a de chance d'être dans la posture couchée (hypothèse H_4). Ce modèle permet de distinguer les postures deux à deux les plus semblables.
- un modèle pour r_3 *et* r_4 : l'idée de ce modèle est que la position la plus compacte est la position accroupie (hypothèse H_3). Ce modèle permet de faire ressortir la position *accroupie* par rapport aux autres postures sachant que la plus grosse difficulté est de distinguer *assis* et *accroupi*.

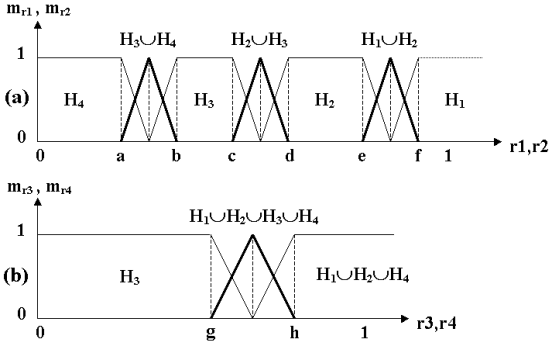

Figure 31 : modélisation des mesures et masses d'évidence élémentaire associées

Sur la Figure 31, les valeurs *a, b, c,..., l* sont des seuils qui ont été déterminés grâce à un ensemble de statistiques sur l'évolution des mesures r_i pour chacune des postures d'une base d'apprentissage.
Avec les modèles choisis, il est possible d'associer une masse d'évidence élémentaire à chaque mesure disponible. La combinaison de ces masses d'évidence par la théorie de l'évidence et la loi de combinaison de Dempster (cf. § 5.2.2) permet d'obtenir une masse d'évidence associée à chaque posture ou combinaison de postures. On retient la posture la plus probable.

Résultats

Nous avons évalué les taux de classification d'une part sur la base dite d'apprentissage (à savoir celle utilisée pour l'apprentissage des seuils de modélisation des mesures cf. Figure 31) et d'autre part sur une base de test. Les Tableau 5 et Tableau 6 fournissent les taux de classification obtenus dans chaque cas. En colonne, nous avons la posture réelle et en ligne la posture reconnue par le système. Dans le cas du Tableau 5, les résultats sont bons voire très bons (taux moyen de reconnaissance de 88,2%) exceptés pour la posture *accroupi*. Ceci vient du fait qu'il existe une réelle variabilité inter personne dans la manière d'être accroupi (mains touchant le sol, mains sur les genoux…) ce qui fait que, pour certaines images, cette posture est classée dans la catégorie des postures *inconnues* H_0. Dans le cas du Tableau 6, les taux de reconnaissance sont un peu moins bons car pour ces séquences test, aucune indication particulière n'a été donnée aux personnes pour réaliser chacune des postures cibles. Les personnes pouvaient se déplacer et exécuter librement chaque posture d'où une variabilité beaucoup plus grande. En particulier, le cas d'une personne assise avec les bras en l'air n'a pas été envisagé dans la phase d'apprentissage. Le taux moyen de reconnaissance sur la base de test est tout de même de 80,8%.

Syst / H	H_1	H_2	H_3	H_4
H_0	0	11,1	30,5	5,5
H_1	100	0	0	0
H_2	0	88,9	0	0
H_3	0	0	69,5	0
H_4	0	0	0	94,5

Tableau 5 : matrice de confusion pour la base d'apprentissage

Syst / H	H_1	H_2	H_3	H_4
H_0	0,3	16,5	31,7	0
H_1	99,7	0	0	0
H_2	0	79,8	24,7	0
H_3	0	3,7	43,6	0
H_4	0	0	0	100

Tableau 6 : matrice de confusion pour la base de test.

7.3 Ce qu'il reste à faire

Imaginons que l'on modifie légèrement les 4 postures considérées de la manière suivante :

- *assis => s'asseoir* ;
- *debout => se lever* ;
- *couché => tomber* ;
- *accroupi => s'accroupir*

on voit alors apparaître la notion de mouvement et donc la notion de posture dynamique. L'idée de la suite de ces travaux est donc de modifier le système proposé afin de conduire à la reconnaissance de postures non plus uniquement statiques mais de postures intrinsèquement dynamiques. De même que pour la reconnaissance dynamique des expressions faciales, le travail va consister à ajouter des données dynamiques (données de vitesse par exemple) et à les fusionner dans un système utilisant une mise en œuvre dynamique de la théorie de l'évidence.
Un autre point qui semble intéressant sera de réfléchir au développement d'une stratégie d'analyse multi-échelle de « gestes » humains. A partir d'une caméra en champ large qui s'intéresse à l'analyse du comportement d'une personne dans son ensemble, comment déclencher (ou non) et asservir une seconde caméra afin de la focaliser sur le visage de la personne et de faire une analyse plus fine de ce qui se passe sur son visage (expressions faciales, mouvements de tête…).

8 Classification de gestes de la main

L'analyse automatique de gestes de la main à partir de la vidéo est directement associée à l'idée de développer des interfaces homme machine plus simples et plus intuitives. Par exemple, il est naturel de décrire un large espace en pointant chacune de ses parties avec l'index. Plus généralement, dans la vie de tous les jours, les gestes de la main sont utilisés à des fins de manipulations d'objets et à des fins de communication. Afin de reconnaître automatiquement un geste de la main, la démarche la plus courante consiste d'abord à localiser la zone d'intérêt de la main puis à faire un choix de modèle de main (soit un modèle 3D, soit un modèle d'apparence) et à estimer les paramètres de ce modèle par extraction d'informations pertinentes dans l'image (contours, forme de la main, orientation, localisation de la main par rapport au reste du corps…) et enfin à procéder à la phase de reconnaissance ou de classification du geste. Remarquons que chacune de ces étapes est un problème à part entière. Par exemple, la segmentation de la zone d'intérêt de la main a conduit aux développements de très nombreuses méthodes [Ong05]. En ce qui concerne la phase de reconnaissance du geste, la démarche consiste soit à reconnaître le geste dans sa globalité soit à décomposer le geste en un ensemble de « sous-gestes » à reconnaître simultanément (par exemple, association d'un mouvement global de la main et du mouvement spécifique de certains doigts). Les méthodes les plus couramment utilisées pour la reconnaissance de gestes de la main sont des classifieurs à base de réseaux de neurones, à base de HMM ou encore à base d'analyse en composantes principales ou d'analyse linéaire discriminante.
Une autre caractéristique importante d'un geste de la main est qu'il s'agit d'un geste intrinsèquement dynamique. La prise en compte d'informations relatives à la trajectoire et au mouvement de la main doit donc être effective dans la mesure du possible.
Le lecteur pourra se référer aux deux articles de synthèse [Pavlovic97, Ong05] pour plus de détails sur les méthodes d'analyse et d'interprétation de gestes de la main. Remarquons que même si l'article [Ong05] s'intéresse exclusivement au langage des signes pour mal-entendants, la plupart des méthodes exposées sont génériques et peuvent s'appliquer à tout autre geste de la main.
Pour notre part, nous nous intéressons au problème de la reconnaissance de gestes de la main particuliers à savoir ceux associés à un langage destiné aux personnes mal-entendantes : le Langage Parlé Complété. Notons que ce langage est différent de la langue des signes en présentant entre autre la particularité d'être proche du langage oral (même structuration de phrases) mais aussi c'est un langage beaucoup plus simple à reconnaître automatiquement (moins de gestes différents, gestes quasi plans et localisés dans l'espace).

Les travaux évoqués dans cette partie sont en cours de développement dans le cadre de la thèse Cifre France Télécom R&D de Thomas Burger dont la soutenance est prévue pour l'automne 2007.

8.1 Introduction au LPC

Les moyens de communication utilisés par les personnes auditivement handicapées sont divers et peuvent être découpés en trois catégories :

- Les personnes malentendantes : elles utilisent la lecture labio-faciale en complément de la voix. Les personnes appareillées avec une prothèse auditive exploitent les deux canaux, visuel et sonore. Même si le canal sonore est très pauvre, la personne a des repères auditifs.
- Les personnes sourdes profondes et oralistes : elles s'appuient sur la lecture labiale avec ou sans l'aide du Langage Parlé Complété (LPC) qui est décrit par la suite.
- Les personnes sourdes gestualistes : elles utilisent la Langue des Signes.

Ici nous nous intéressons au deuxième cas et plus particulièrement à la reconnaissance automatique des gestes du LPC. Ce langage, développée par Cornett en 1967 pour l'Anglo-Américain, utilise un jeu de clés manuelles présentées près du visage du locuteur pour venir compléter la lecture labiale (cf. Figure 32 et Figure 33), cette dernière étant par nature incomplète et confuse (une même image labiale pouvant être associée à plusieurs phonèmes, par exemple « u » et « ou »). La position de la main (parmi cinq positions possibles) code la voyelle, la forme de la main (huit clés digitales) code la consonne. Ainsi une position et une forme de main données codent simultanément la consonne C et la voyelle V d'une séquence CV. Une même clé digitale est utilisée pour coder des consonnes possédant un contraste visuel important, et inversement les consonnes ayant des formes aux lèvres très proches sont codées par deux formes de main très différentes. Il en est de même pour le regroupement des voyelles visuellement contrastées à l'intérieur des cinq positions possibles. C'est donc la complémentarité des deux informations visuelles des lèvres et de la main qui, par intersection, permet de désambiguïser la parole visuelle.

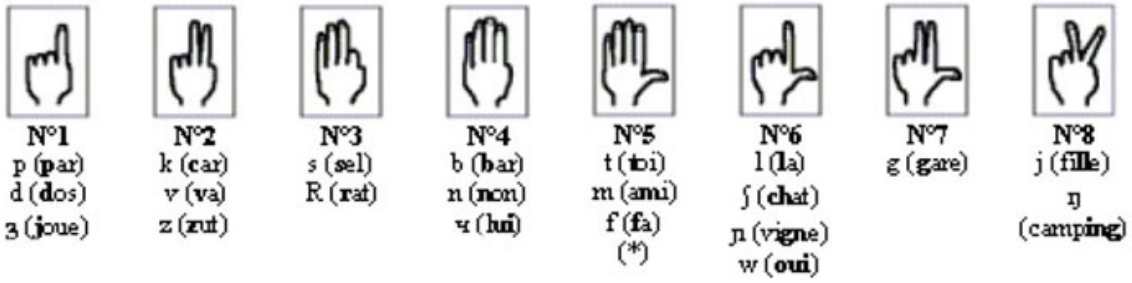

Figure 32 : huit configurations possibles pour la main (codage d'une consonne)

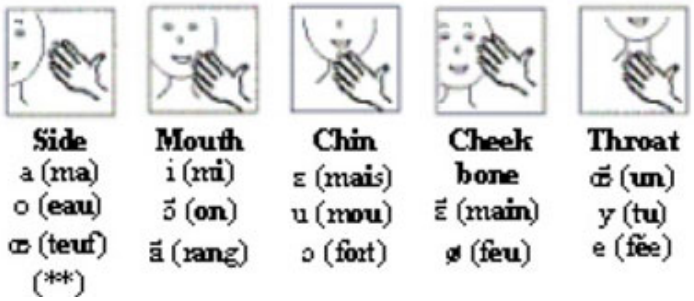

Figure 33 : cinq positions possibles par rapport au visage (codage d'une voyelle)

L'objectif est de développer un système temps réel de classification des gestes du LPC (configuration + position). Afin de rendre la segmentation de la main par rapport au visage

plus aisée, nous ajoutons ici l'hypothèse que le codeur porte un gant dont la couleur est significativement différente de celle de la peau (cf. Figure 34). Cette contrainte a été discutée avec des codeurs professionnels qui ne s'y sont pas opposés.

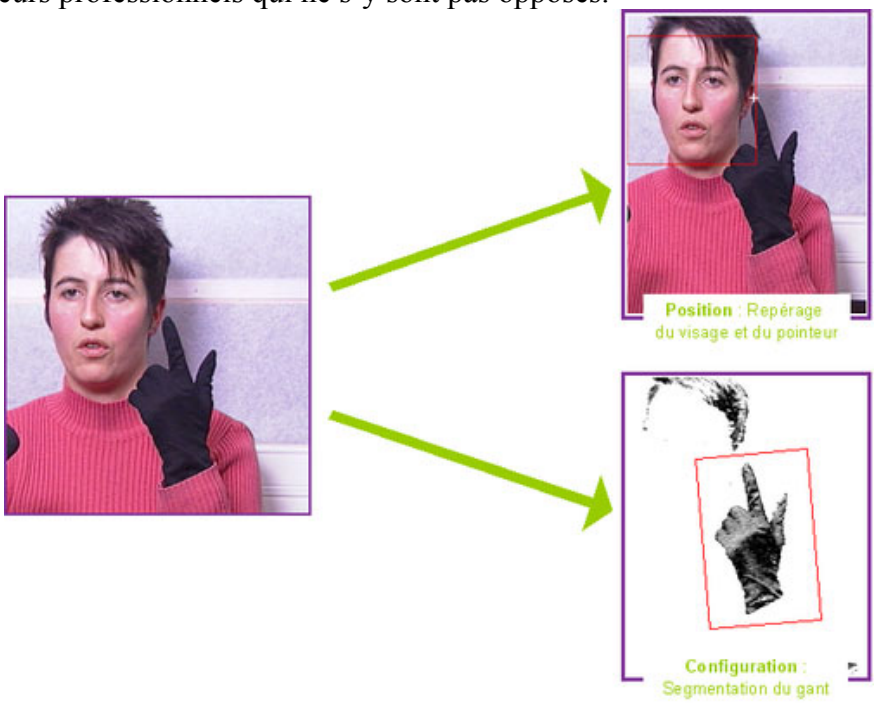

Figure 34 : à gauche, une image de séquence de LPC (config 6 + position côté) ; à droite, extraction du pointeur pour la reconnaissance de la position et segmentation de la main pour la reconnaissance de la configuration.

8.2 Travail préliminaire : données 2D ou données 3D ?

Dans le cadre d'un travail préliminaire réalisé en collaboration avec le laboratoire ITI_CERTH de Thessalonique (partenaire du réseau d'excellence Similar), nous nous sommes posés la question de savoir si la reconnaissance des gestes du LPC nécessite une prise de vue monoculaire ou stéréo. Ce travail a conduit à la publication **[CI_Caplier04]**.

Dans une première approche, les images proviennent d'une caméra numérique monoculaire, ses images étant segmentées par seuillage après apprentissage de la couleur du gant. De plus, les conditions d'éclairage ont été ajustées afin de rendre la segmentation la plus simple possible (cf. Figure 35). Notons toutefois que les conditions d'acquisition ne sont pas du tout réalistes (très forte lumière).

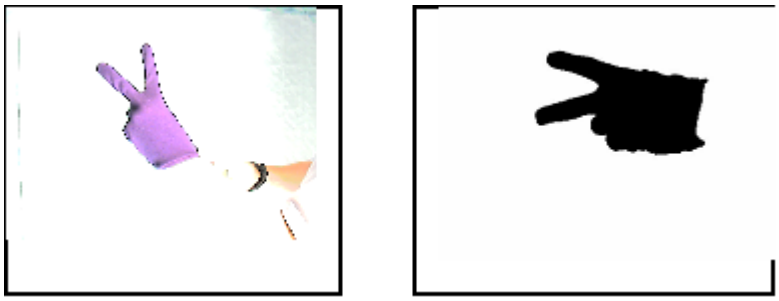

Figure 35 : à gauche, image acquise avec un fort éclairement ; à droite, image segmentée après simple seuillage

Dans une seconde approche, les images ont été acquises avec une caméra capable de faire des acquisitions 3D couleur. Le fonctionnement de la caméra utilisée est basé sur le principe de triangulation active, utilisant une version améliorée et étendue de l'approche Coded Light pour l'acquisition de données 3D [Forster01]. L'opération de base de cette caméra est la projection de patterns de couleur sur une scène suivie de la mesure des déformations de cette

lumière à la surface des objets éclairés. De ce fait, il est possible d'acquérir une image couleur ainsi qu'une carte de profondeur. La segmentation de la main utilise alors les informations de profondeur. La segmentation obtenue est de moins bonne qualité que celle obtenue dans le cas 2D mais les conditions d'acquisition sont réalistes.

Figure 36 : à gauche, images couleur ; au milieu, cartes de profondeur ; à droite, mains segmentées.

Sur chacun des masques de main segmentée, un ensemble d'attributs invariant aux translations, aux rotations et aux homothéties a été calculé. Le choix s'est porté sur les invariants de moment de Hu. Ces invariants sont ensuite les données d'entrée d'un perceptron multi-couche. La comparaison des classifications portant sur des données issues de la segmentation 2D d'une part et portant d'autre part sur les données issues de la segmentation 3D montrent que :

- dans le cas où la segmentation 2D est de très bonne qualité, la classification obtenue par la suite est proche de 100% de bonne reconnaissance pour toutes les configurations.
- les invariants de moment de Hu sont discriminants.
- en cas de mauvaise segmentation des images 2D, il est préférable d'utiliser des données 3D.

Tout le challenge consiste alors à développer une méthode de segmentation de la main qui puisse fournir des résultats de bonne qualité sur les images 2D acquises dans des conditions d'éclairement réalistes.

8.3 Segmentation de la main

Ce travail a conduit à la réalisation d'un premier démonstrateur de segmentation et de classification des gestes du LPC qui a été décrit en détails dans [**CI_Burger05**].

8.3.1 Approche orientée région : apprentissage des caractéristiques du gant et seuillage.

L'apprentissage de la couleur du gant a lieu sur la première image acquise. A partir d'une zone de gant sélectionnée manuellement par l'utilisateur dans la phase d'initialisation, on effectue une modélisation par des gaussiennes des distributions de luminance et de chrominance des pixels associés au gant (cf. Figure 37).

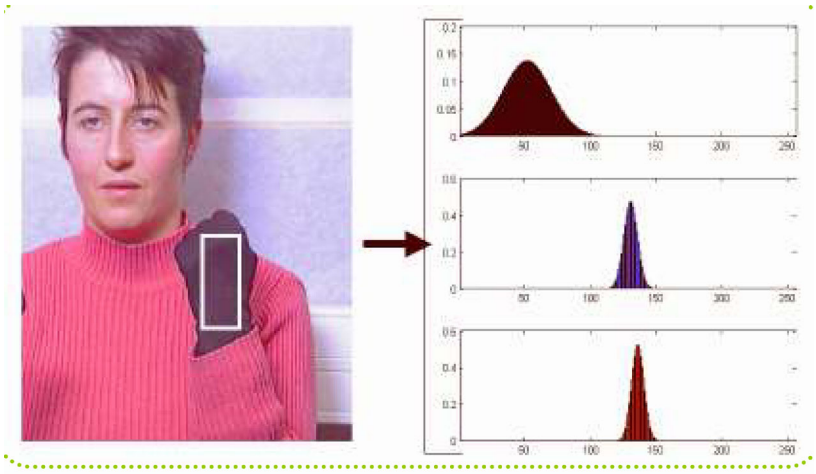

Figure 37 : apprentissage des caractéristiques du gant : à gauche, sélection manuelle de la zone de gant ; à droite, histogramme de luminance et de chrominances des pixels du gant.

Une segmentation des pixels appartenant au gant est effectuée par seuillage sur ces différentes gaussiennes. Le choix du seuil est contrôlé par l'utilisateur lors de la phase d'initialisation du système. La Figure 38 présente quelques résultats de segmentation ainsi obtenus.

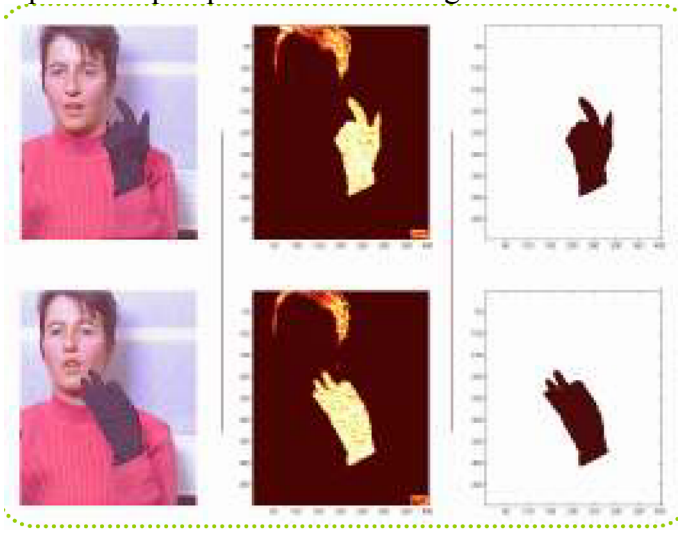

Figure 38 : de gauche à droite, image de main, image de probabilité d'appartenance au gant ; masque de la main segmentée après seuillage sur la probabilité.

8.3.2 Approche orientée contour : optimisation d'un contour actif

Afin d'affiner la segmentation de la main, une approche orientée contour est envisagée. Elle est dérivée de la méthode de segmentation par modèle de formes actif décrite dans [Chesnaud00] et inspirée des travaux initiaux de Cootes sur les *Active Shape Models* [Cootes95]. L'idée est de définir un contour comme la ligne de séparation entre deux régions de caractéristiques homogènes. Le contour recherché est défini par un ensemble de points qui sont déplacés à tour de rôle et dont le déplacement est validé pour peu qu'il tende à homogénéiser les caractéristiques de luminance et/ou de chrominance des deux zones (à l'intérieur et à l'extérieur du contour). La Figure 39 présente un exemple de contour de main obtenu par optimisation et déformation d'un *ASM*. Bien que conduisant à une segmentation plus précise que celle obtenue sur un critère orienté région, cette méthode n'est cependant pas capable de faire ressortir le contour présent entre deux doigts voisins lorsque ces deux doigts se touchent. La segmentation orientée contour ne permet donc pas un comptage plus précis des doigts que la segmentation orientée région.

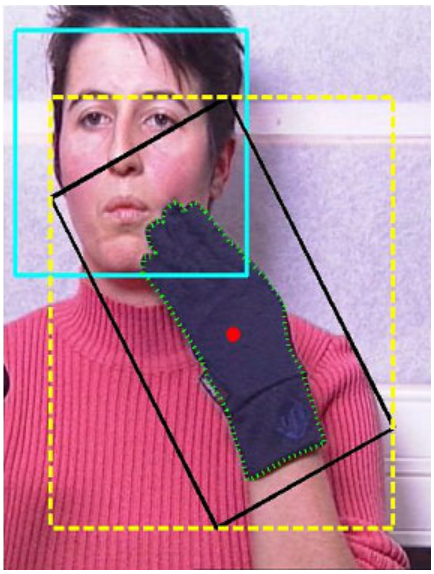

Figure 39 : en pointillés verts, contour de la main obtenu par la méthode des ASM.

8.4 Reconnaissance des configurations du LPC

8.4.1 Sélection des images cibles

Une séquence de gestes du LPC est une séquence pour laquelle la main est constamment en mouvement de manière à passer continûment du codage d'une syllabe à une autre. De ce fait, une séquence d'images de gestes de LPC est constituée d'un ensemble d'images correspondant à des configurations dites cibles (à savoir des couples (forme de main + position)) pouvant être associées à une ou plusieurs syllabes mais elle contient aussi un grand nombre d'images correspondant à des transitions entre deux configurations cibles. Le traitement des images de transition est a priori inutile puisque les formes et les positions intermédiaires de la main ne sont associées à aucun message particulier. D'où l'idée d'essayer de détecter a priori les images de la séquence qui correspondent à des configurations cibles afin de ne traiter que ces images d'intérêt. Afin de détecter les « bonnes » images ou images cibles, l'idée est d'appliquer sur la séquence des boites englobant la main le détecteur de mouvement biologique décrit au paragraphe 6.1. En effet, rappelons que le détecteur biologique de mouvement est tel qu'il présente un minimum d'énergie à la sortie de l'IPL en cas d'arrêt ou de ralentissement du mouvement présent dans la zone d'intérêt. Or dans le cas des gestes du LPC, chaque configuration cible est maintenue quelques instants (même infimes) si bien qu'il est possible de détecter les images correspondant aux configurations cibles en détectant les minima d'énergie présents à la sortie de l'IPL. La Figure 40 montre un exemple d'évolution de l'énergie en sortie de l'IPL pour une séquence de gestes de LPC correspondant au codage de la phrase « nous traquions bien Euler pendant son footing urbain ». Le détecteur a permis de faire ressortir 23 images cibles à traiter. Remarquons que sur cette séquence, il existe une zone durant laquelle la codeuse a fait une pause avant de reprendre la phrase, cette zone n'a pas été considérée ici.
La détection préalable des images à configuration cible présente un grand intérêt dans notre optique de reconnaissance temps réel des gestes du LPC. En effet, elle permet de réduire le nombre d'images à traiter. Par ailleurs, la détection des instants associés aux image cibles a aussi un autre intérêt, celui de pouvoir étudier le synchronisme entre les gestes de la main et le mouvement des lèvres. Il s'avère en effet que la main est toujours en avance sur les lèvres mais que cette avance n'est pas constante au cours du temps.

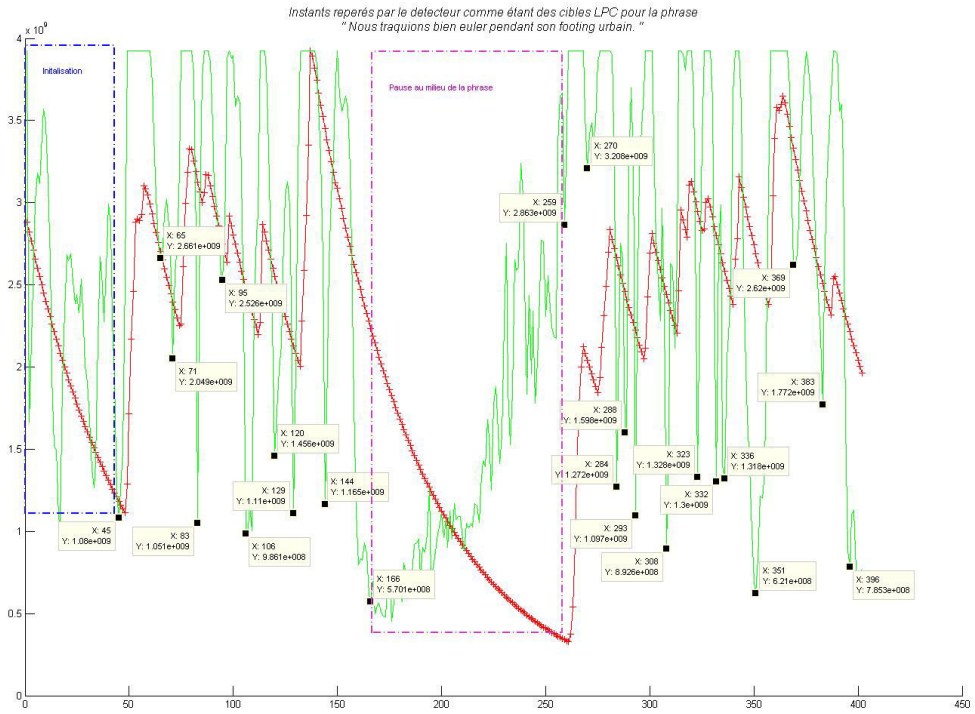

Figure 40 : en vert, évolution temporelle de l'énergie à la sortie de l'IPL. Sur cette courbe, les carrés noirs correspondent aux images à configuration cible.

8.4.2 Principe du système de classification : définition d'un codage numérique associé à chaque configuration

L'expertise que nous avons sur l'analyse des séquences de gestes du LPC nous a amené à définir une stratégie de reconnaissance de chacune des configurations de la main en dénombrant le nombre de doigts présents et en associant un code numérique à chacune des configurations en fonction du nombre de doigts présents. Afin d'obtenir des codes numériques différents pour la plupart des configurations, nous avons associé un poids de 0.5 au pouce et un poids de 1 à tous les autres doigts. Nous obtenons ainsi la codification de la Figure 41 pour chacune des configurations. La configuration 1 est codée par la valeur 1 car elle ne présente qu'un seul doigt différent du pouce, la configuration 7 est codée par la valeur 2.5 car elle présente deux doigts et le pouce. L'examen du codage proposé montre qu'il est possible de différentier toutes les configurations à l'exception des configurations 2 et 8 qui présentent toutes les deux deux doigts (autres que le pouce), la seule différence résidant dans l'écartement de ces deux doigts. Nous pensons que la différentiation de ces deux configurations pourra se faire en analysant justement l'écartement entre les deux doigts présents.

V_{ref}	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
HS		X						X		

Figure 41 : codage numérique associé à chaque configuration

8.4.3 Information de bas niveau : nombre de doigts, grandeurs caractéristiques

Le principe global de classification étant assez simple, toute la difficulté réside dans l'extraction et la fusion d'informations permettant d'en déduire le nombre de doigts présents sur une image de main donnée. La démarche proposée à l'heure actuelle consiste à rechercher une modélisation de la main sous la forme proposée sur la Figure 42-gauche. Cette modélisation requiert la détermination du poignet, de la paume et des doigts présents.

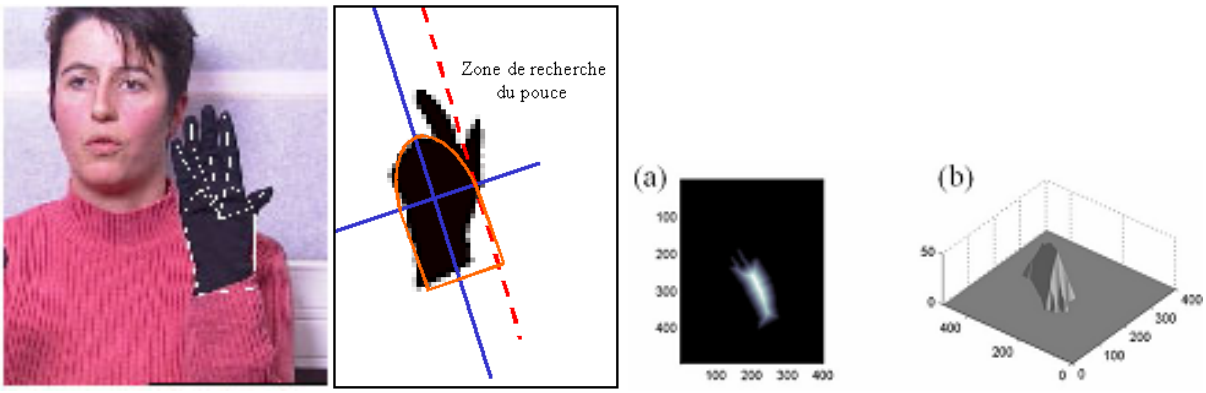

Figure 42 : à gauche, modèle de la main ; au milieu, masque de la main avec axes d'inertie, paume, poignet et zone de recherche du pouce ; à droite, visualisation 2D et 3D de la carte de transformée de distance.

A partir du masque de la main issu de l'étape de segmentation sur la base des caractéristiques de couleur du gant, des paramètres globaux de surface, moments et axes d'inertie permettent d'en déduire entre autre l'orientation de la main puis la position du poignet qui est a priori située dans la région la plus « basse » du masque segmenté (voir Figure 42-milieu).
Sur le masque de la main, une transformation de distance est appliquée. Elle est calculée de la manière suivante :

$$dist\_transform(p) = d(p, contour)$$

En chaque point p de la carte de distance, on calcule la distance de ce point au point le plus proche appartenant au contour du masque de la main. Un exemple de carte de distance obtenue est donné sur la Figure 42-droite.
Cette transformation présente l'avantage de pouvoir localiser aisément le centre de la main qui correspond au point pour lequel la distance au contour est maximale. Ce point est supposé être le centre de la paume. A partir de ce point, il est possible d'en déduire un demi cercle modélisant la paume de la main (cf. Figure 42-milieu).
Pour la détection des doigts potentiels, nous nous limitons à l'étude des pixels du masque de la main situé au dessus de la limite circulaire de la paume. Nous recherchons alors les zones de pulpes de doigts potentiels en élaborant deux cartes de probabilité :

- une carte de probabilité d'un pixel d'appartenir à un doigt, cette carte est basée sur l'idée que les pixels des doigts sont localisés sur les bords du masque de la main (pixels pour lesquels la valeur dans la carte de distance transformée est faible).
- une carte de probabilité qu'un pixel du contour du masque de la main appartienne à un cercle, cette carte repose sur l'idée que le bout d'un doigt engendre un contour de type circulaire (utilisation de la transformée de Hough).

La combinaison de ces deux cartes de probabilité dont on va extraire les points de probabilité combinée la plus élevée conduit à la sélection d'un ensemble de points doigts candidats (voir Figure 43-droite). Comme nous ne savons a priori pas le nombre de doigts présents dans la combinaison courante, nous devons considérer au minimum les 5 points les plus probables. Malheureusement, il s'avère qu'en ne considérant ces 5 points seulement, il arrive que l'on obtienne certaines détections parasites et que l'on perde un ou deux doigts. En revanche, une

analyse plus précise des résultats obtenus montre qu'en considérant les 10 points les plus probables, tous les doigts sont toujours présents. Ils s'agit alors de faire la distinction entre les points appartenant réellement à un doigt et les détections parasites. Ceci nécessite la prise en compte d'informations supplémentaires et est en cours de développement.

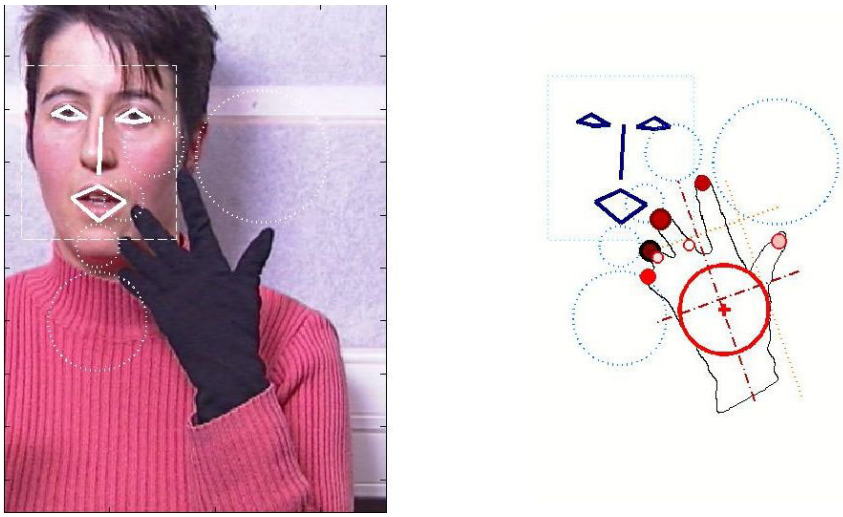

Figure 43 : à gauche, image à traiter ; à droite, détection de points susceptibles d'appartenir à un doigt (cf. cercles rosés)

8.5 Ce qu'il reste à faire

Les travaux sur la reconnaissance des gestes du LPC sont des travaux récents. Au stade actuel, il va falloir dans un premier temps développer une méthode d'identification précise des doigts présents sur une image donnée. Ceci se fera dans la continuité de ce qui a été présenté dans la partie 8.4.

Une deuxième phase importante va consister en la reconnaissance de la seconde information associée aux gestes du LPC à savoir l'identification de la position de la main par rapport au visage. Pour ce faire, nous nous appuierons bien évidemment sur les travaux que nous avons déjà développés sur la localisation du visage, des yeux, des sourcils et de la bouche.

Un dernier point à aborder et non des moindres sera de réfléchir à une architecture adaptée aux algorithmes qui auront été développés afin de pouvoir développer un prototype de reconnaissance des gestes du LPC fonctionnant en temps réel.

9 Description des projets de recherche

Ce chapitre est destiné à montrer comment les recherches que nous avons développées ont été utilisées dans des projets et des applications spécifiques.

9.1 Le projet européen Art-live (prototype d'architecture et d'outils-auteurs pour flux d'images en temps-réel et nouvelles expériences vidéo). (01/01/00 à 01/04/02)

Le but d'*art.live* était de développer une architecture et un ensemble d'outils, à la fois génériques et orientés application, pour l'amélioration des espaces narratifs. Dans ce but, *art.live* rassemblait des ingénieurs en traitement de signal (le laboratoire TELE de l'Université Catholique de Louvain, Le LIS, le laboratoire de traitement du signal de l'Ecole Polytechnique de Lausanne), des chercheurs en informatique (TIMC, ADETTI de Lisbonne), des industriels (ADERSA, FASTCOM) et des auteurs multimédia (CASTERMAN). Le projet a permis de réaliser deux essais grandeur nature et d'établir de nombreuses démonstrations (IST2000, ICAV3D…).
L'objet du projet était de développer des environnements mélangeant le monde réel et le monde virtuel pour des applications multimedia. Il en résulte la création de scènes de réalité mixtes pour lesquelles des personnages filmés sont extraits de leur environnement réel afin d'être replacés dans un environnement virtuel dans lequel il sera possible d'interagir. Par exemple, sur la Figure 44, le personnage extrait est placé dans un décor virtuel dans lequel il lui est demandé d'attraper tous les papillons d'une couleur donnée, ceci ayant pour effet de le transformer en papillon. Des scénarios plus complexes faisant intervenir plusieurs personnages filmés par des caméras différentes ont été également envisagés et réalisés.

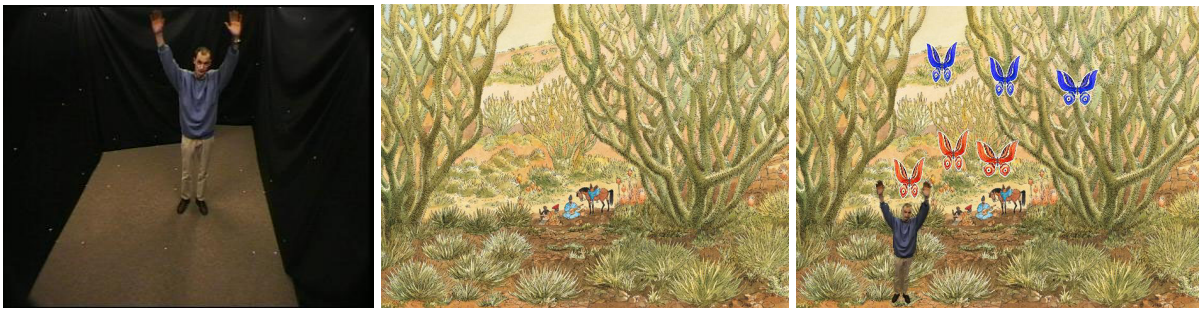

Figure 44 : projet artlive : à gauche, personnage en mouvement dans une scène réelle ; au milieu, fond virtuel ; à droite, incrustation du personnage ainsi que d'objets virtuels (papillons) dans le décor virtuel

La contribution du LIS à ce projet a porté sur l'analyse des scènes réelles en vue de l'extraction et du suivi de personnes ainsi que de leurs mains. En effet, la connaissance à chaque instant de la position de chaque main est nécessaire pour déclencher par la suite des événements particuliers dans la scène de réalité mixte.
Les algorithmes proposés dans le cadre de ce projet ont été développés dans la suite de ce qui a été proposé dans le cadre de ma thèse et dans le cadre de la thèse de Vincent Girondel (cf. §7.1.) Ont été proposés des algorithmes :

- de détection de personnes en mouvement (cf. **[CI_Caplier01a, CN_Bonnaud01]).**
- de suivi de personnes (cf**. [CI_Girondel04])**
- de détection et de suivi de la tête et des mains d'une personne (cf. **[CI_Girondel02]**)

Ce projet s'est accompagné de deux démonstrateurs présentés et testés lors de deux manifestations grand public :

- lors du festival « Les jardins et la bande dessinée » (cf. Figure 45) organisé par la marie de Paris et ayant eu lieu entre septembre 2000 et mars 2001 dans les jardins de Bercy. Le scénario appelé "Adèle@Bercy" consistait à proposer aux visiteurs d'aider l'héroïne Adèle Blansec de Jasques Tardi publié chez Casterman à résoudre certains mystères dans Paris. Neuf enquêtes ont été proposées, chacune d'entre elle correspondant à un endroit différent sur une carte de Paris (cf. Figure 46). A chaque fois était proposée une succession de scènes qui mélangeaient le monde réel (plusieurs caméras étaient mises en place dans les jardins) et le monde virtuel. Chaque joueur était invité à traquer les méchants robots et à les détruire en cliquant dessus ou alors à retrouver leurs amis dans les personnages extraits des scènes réelles filmées.

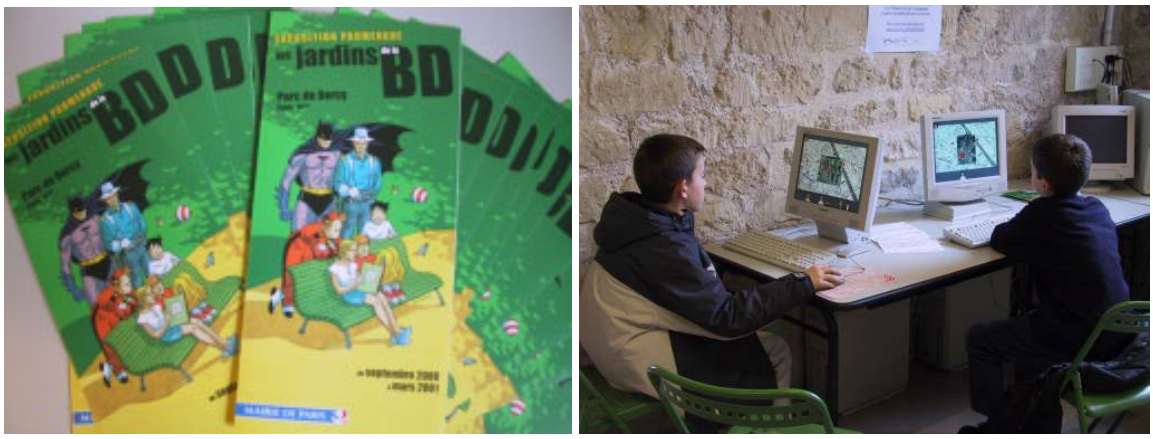

Figure 45 : à gauche, plaquette de présentation de l'exposition ; à droite, exemples d'enfants testant le jeu proposé.

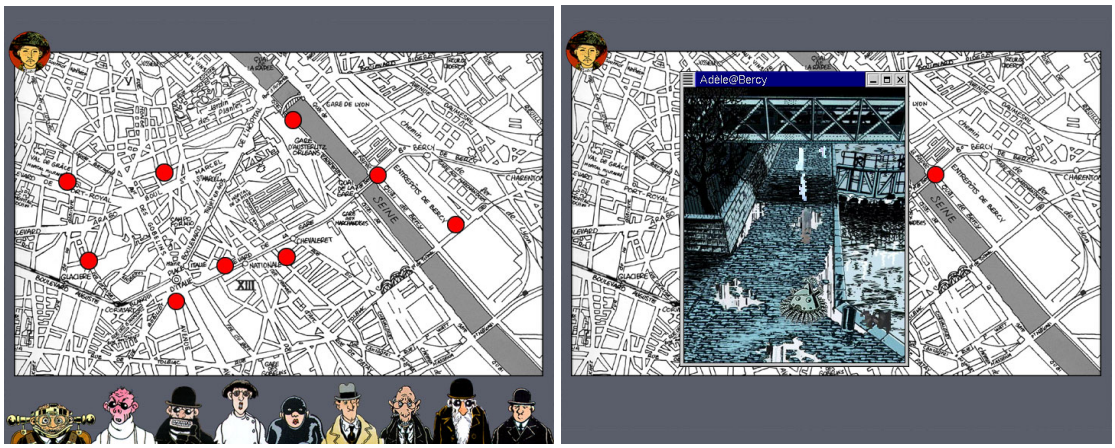

Figure 46 : à gauche, plan de Paris ; à droite, scène de réalité mixte du jeu.

- le second démonstrateur a été installé pendant deux semaines en novembre 2001 dans les locaux de la Saline Royale à Arc et Senan, France. Basé sur le paradigme du miroir magique, le ou les scénarios faisaient intervenir deux joueurs simultanément confrontés à 6 jeux différents dont ils devaient être acteurs. Les Figure 47 et Figure 48 présentent quelques illustrations des jeux proposés. Sur la Figure 47, chaque personnage situé devant l'écran doit toucher une à une l'ensemble des pièces du puzzle afin de le reconstituer. Sur la Figure 48, il s'agit d'une réplique du jeu du pendu pour laquelle chaque joueur propose une lettre à tour de rôle. Chaque lettre choisie

doit être attrapée par le joueur et déplacée à la bonne place. Ces petits jeux ont été testés par des enfants lors de sortie scolaire.

The Wallawa River

Figure 47 : puzzle géant (images et dessins de © Casterman)

The City of Vertigo

Figure 48 : jeu du pendu (images et dessins de © Casterman)

A l'issue de chacun des deux tests grandeur nature, une évaluation auprès du public a été menée. Il ressort que ce type de système est bien apprécié du public même si plusieurs étapes comme la segmentation des personnages peut s'avérer parfois imparfaite.
Plusieurs publications communes à l'ensemble des partenaires du consortium Artlive ont été faites (cf. **[CI_Artlive01, CI_Artlive02]**).

9.2 Le projet RNRT Tempo-Valse : Terminal Expérimental MPEG-4 Portable de Visiophonie et Animation Labiale ScalablE (01/01/00 à 31/12/02)

Dans la perspective des futurs réseaux mobiles 3G et au-delà, le projet TempoValse s'était fixé de réaliser une maquette de terminal multimédia portable basé sur la normalisation MPEG-4.
Les prévisions d'accroissement des débits ouvrant de nouvelles possibilités en matière d'applications multimédia, le projet visait à réaliser l'étude de faisabilité d'un terminal portable multimédia basé MPEG-4, dédié à la consultation et la communication interactive multimodale (son + visuel). L'application visée était de qualité échelonnable allant de la vidéo à l'animation de visages parlant. La maquette devait permettre d'étudier l'amélioration de l'intelligibilité par l'apport de la modalité visuelle dans le cas de communication en milieu bruyant (mobiles). Il fallait évaluer aussi l'acceptabilité sociale de ce type de terminal avec microphone, caméra et afficheur.
Un effort particulier a été mené en vue du développement des briques technologiques et de l'architecture du terminal en vue d'expérimenter une application de visiophonie et d'animation labiale, « service prétexte » de communication multimodale scalable.
Les grandes lignes du projet peuvent se résumer aux thèmes suivants :

- définition d'un terminal cible avec les interfaces homme machine appropriées aux contraintes de mobilité et étude sur l'ergonomie, l'usage et l'apport de la modalité visuelle.

- études technologiques logicielles et matérielles concernant la chaîne de traitement des flux audio, vidéo, animation de visages parlants et flux MPEG-4.
- démonstration de visiophonie et animation labiale temps réel.

Ce projet a été mené en collaboration avec plusieurs partenaires universitaires : le LIS, l'Institut de la Communication Parlée, l'ENST Paris et plusieurs partenaires industriels : le CNET, ST Micro-électronics, LEP Phillips, Ganimédia.
Au LIS, nous avons principalement travaillé à l'extraction du contour des lèvres en vue de sa transmission sous forme compacte afin d'augmenter l'intelligibilité d'un signal audio en environnement bruité (cf. effet Mc Gurk). Après l'extraction fine et précise du contours des lèvres (cf. travail mené dans le cadre de la thèse de Nicolas Eveno (cf. §5.1), voir les publications **[CI_Eveno01, CI_Eveno02a, CI_Eveno02b, CI_Eveno03, R_Eveno04]**), des paramètres sont extraits afin de rendre compte de l'évolution temporelle de la forme des lèvres au cours de l'élocution. Dans le cadre de ce projet, le système d'acquisition des images est imposé : la caméra est montée sur un casque si bien que le cadrage du visage est fixe (cf. Figure 49).

Figure 49 : à gauche : micro-caméra ; à droite : exemple de séquences d'images acquises par la micro-caméra.

9.3 Le projet Telma (début janvier 2004)

L'objectif du projet TELMA est le développement d'un terminal de téléphonie à l'usage des malentendants. A l'origine de ce projet se trouvent 3 partenaires universitaires (LIS, ICP, CLIPS équipe GEOD), ce projet ayant été financé par l'INPG dans le cadre de la campagne du BQR 2004. A ces trois partenaires universitaires s'ajoute France Télécom R&D qui intervient à l'heure actuelle dans ce projet uniquement au niveau de l'analyse vidéo par l'intermédiaire du financement de la thèse Cifre de Thomas Burger. Afin de compléter le financement nécessaire à la réalisation complète de ce projet, celui-ci a été soumis au RNTS. Le dossier est en cours d'expertise.
Les schémas de principe du système envisagé sont donnés sur les Figure 50 et Figure 51.

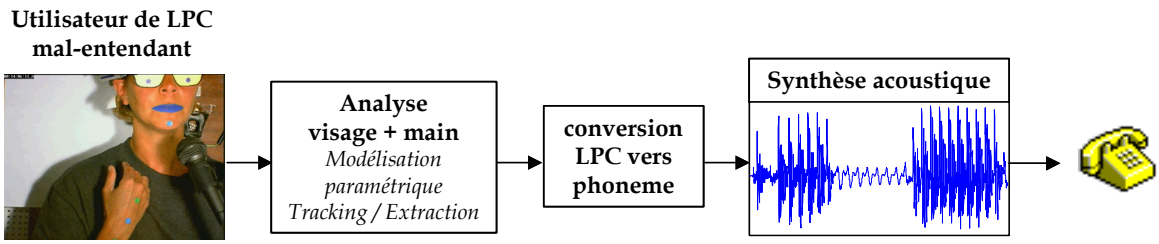

Figure 50 : synthèse acoustique à partir de la vidéo LPC

Sur la Figure 50, on envisage la synthèse du signal de parole à partir de l'analyse conjointe de la forme des lèvres et des gestes du LPC. En effet, l'ensemble des informations portées par les lèvres et les gestes de la main permet aux malentendants de décoder le message de parole.

Nous travaillons sur la partie relative à l'analyse des séquences de LPC tant au niveau de l'extraction des contours des lèvres (thèse de Nicolas Eveno) qu'au niveau de la reconnaissance automatique et temps réel des gestes du LPC (thèse de Thomas Burger).

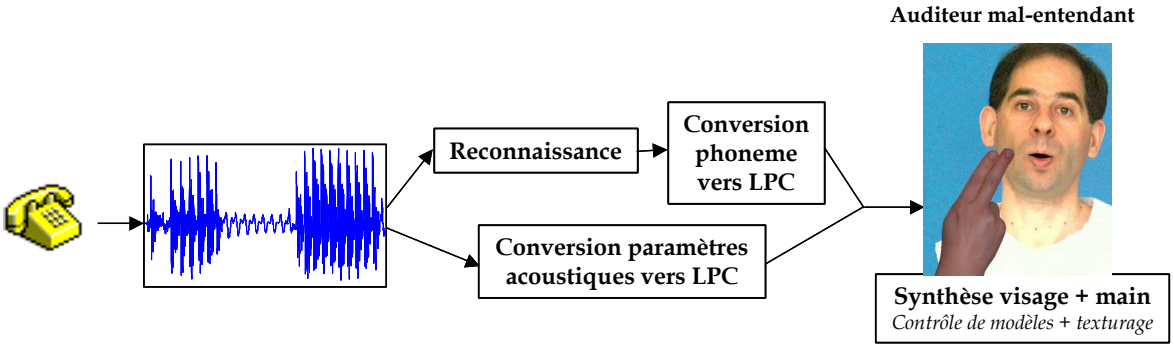

Figure 51 : synthèse vidéo des gestes du LPC à partir de l'analyse et de la reconnaissance audio

Sur la Figure 51, on envisage la synthèse de la vidéo du LPC (mouvement de lèvres + gestes de la main) à partir de l'analyse et de la reconnaissance du signal audio. Ce mode de fonctionnement est pris en charge par l'ICP et le CLIPS (équipe GEOD).

9.4 Le réseau d'excellence Similar (début : décembre 2003)

9.4.1 Description et objectifs

Le réseau d'excellence européen Similar (www.similar.cc) regroupant une trentaine de partenaires a pour but de créer de nouvelles interfaces hommes-machines plus intuitives et donc s'inspirant des processus de communication face à face (cf. Figure 52).

Figure 52 : réseau d'excellence européen Similar

La création d'interfaces homme machine s'inspirant de la communication face à face suppose l'intégration de plusieurs modalités. On distingue les modalités actives telles que la parole, le geste… et les modalités passives telles que l'intonation, la direction du regard, les expressions faciales. Disposant de plusieurs modalités, les enjeux sont multiples :

- comment fusionner toutes ces modalités pour en déduire des actions de haut niveau ?
- toutes les modalités sont elles nécessaires pour tout type d'application ? Faut-il éliminer la redondance entre modalités ou bien l'exploiter ?

La gestion des modalités considérées se fait donc à deux niveaux : le niveau dit de *fusion* qui consiste à combiner entre elles toutes ces informations de nature différente et le niveau dit de *fission* qui consiste à l'inverse à sélectionner au préalable le nombre de modalités optimales pour la communication dans un cadre donné.
La Figure 53 donne une vue synthétique des objectifs du réseau Similar.
Afin de tester les nouvelles interfaces proposées, une plate-forme de tests OPENINTERFACE est développée et enrichie périodiquement (cf. Figure 54).

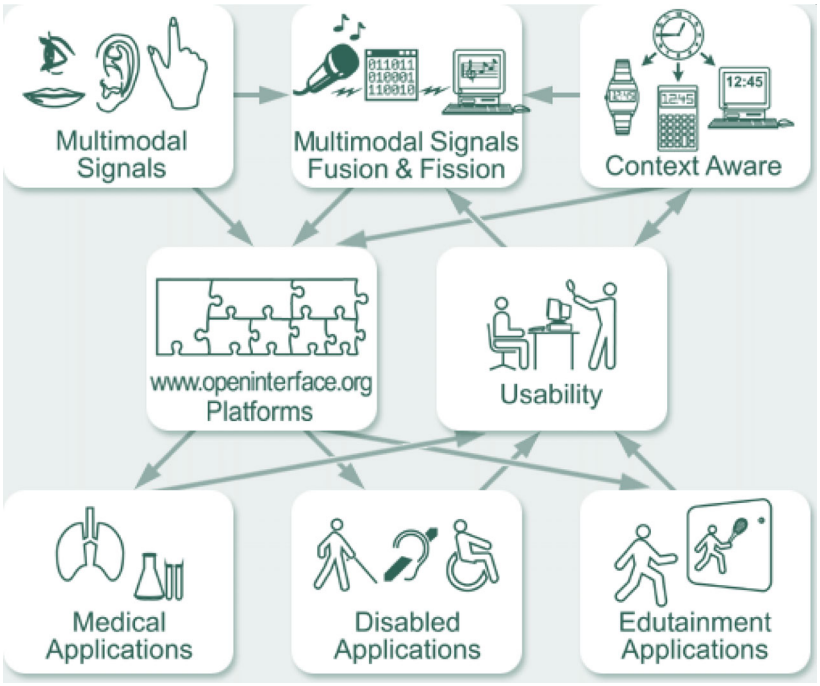

Figure 53 : synthèse des objectifs de Similar.

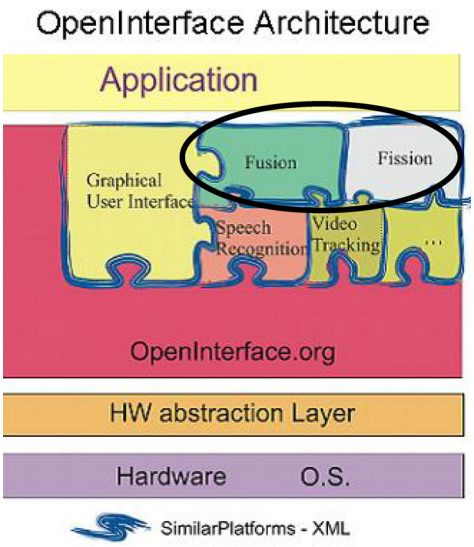

Figure 54 : plate-forme de test OPENINTERFACE

9.4.2 Collaborations et travaux réalisés

L'ensemble des travaux effectués sur l'analyse des mouvements humains s'intègrent dans les objectifs du réseau d'excellence Similar. En effet, l'ensemble des modalités telles que les expressions faciales, la position d'un utilisateur, l'analyse de la direction de son regard, de ses mouvements de tête ou bien encore de certains de ses gestes participe à l'acte de communication en face à face. Toutes ces informations sont donc utiles pour la définition d'interfaces homme machine « intelligentes ».
Plus spécifiquement, nous avons travaillé sur le sous-projet 9 dédié à l'analyse des signaux multimodaux, sur le sous-projet 11 dédié au développement d'interfaces nouvelles pour les handicapés et sur le sous-projet 12 dédié aux applications de type « edutainment » alliant l'apprentissage et le jeu.

9.4.3 Enterface summer workshop : présentation

Une des actions associées au réseau Similar est la tenue chaque année d'une école d'été nommée eNterface (http://www.enterface.net/). L'objectif de cette école d'été est de faire un travail spécifique de recherche et de développement en rassemblant en un même lieu pour une durée de un mois de jeunes chercheurs ainsi que des chercheurs confirmés. L'ensemble des participants est géré sous forme d'équipes attachées à un projet précis. Pour la première édition de ce workshop, sept projets différents ont été retenus dont entre autre :

- Analyse et synthèse de la combinaison geste et parole proposé par l'université d'Istambul, Turquie ;
- Miroir magique multimodal proposé par l'université Catholique de Louvain la Neuve, Belgique ;
- Instrument de musique contrôlé par des modalités biologiques (EEG, EMG…) proposé par l'université de Louvain la Neuve, Belgique ;
- Identification biométrique multimodale proposée par l'université de Crête, Grèce ;
- Interprétation du niveau de vigilance d'un utilisateur de simulateur de conduite proposé par l'université de Louvain la Neuve, Belgique et le Laboratoire des Images et des signaux de Grenoble.

9.4.4 Enterface Project 4 : multimodal focus attention detection in an augmented driver simulator

L'objectif du projet que nous avons proposé et co-encadré est le développement d'un simulateur de conduite capable d'analyser le comportement de son utilisateur. Plus particulièrement, il s'agit de détecter les instants d'hypovigilance du conducteur sur la base de l'analyse d'information vidéo et de détecter également les moments de stress du conducteur sur la base de l'analyse de signaux biologiques tels que l'électrocardiogramme (ECG) et la conductivité de la peau (GSR : Galvanic Skin Response). Dès lors qu'un état d'hypovigilance ou de stress a été détecté, le simulateur envoie des messages d'alarme à l'utilisateur (messages visuels sur le tableau de bord, vibration du volant) que l'utilisateur doit attester avoir reçu (arrêt de la vibration du volant par appui sur un bouton). Un rapport détaillé du travail réalisé a été publié **[CI_Enterface05].**

L'équipe associée à ce projet était composée de 7 personnes venant du LIS, de l'Université Catholique de Louvain la Neuve (Belgique), de la Faculté Polytechnique de Mons (Belgique), de l'Université de Genève (Suisse) et de l'Université de Zagreb (Croatie). En alternance avec Laurent Bonnaud du LIS et Daniela Trevisan de l'UCL, j'ai assuré la direction de ce projet.

La Figure 55 présente une vue globale du démonstrateur réalisé. Donc peut y voir un utilisateur en situation de conduite et sous haute surveillance. En effet, on distingue la présence d'une caméra filmant le visage de l'utilisateur ainsi que la présence de deux électrodes destinées à enregistrer l'ECG de l'utilisateur.

Figure 55 : démonstrateur réalisé

La Figure 56 présente un diagramme de l'architecture associée au simulateur de conduite augmenté. Nous avons opté pour une approche distribuée : un PC fonctionnant sous Windows sert à l'exécution du simulateur de conduite, un PC fonctionnant sous Linux réalise la détection des états d'hypovigilance par analyse de la vidéo du visage de l'utilisateur et un PC fonctionnant sous Windows réalise la détection des états de stress par analyse des signaux biologiques.

Nous allons nous focaliser sur la détection des états d'hypovigilance puisque cette détection résulte de l'utilisation des algorithmes proposés dans la partie 6. En effet, trois signes caractéristiques d'un état d'hypovigilance sont recherchés :

- fermeture des yeux du conducteur pendant plus d'un certain temps ;
- bâillement du conducteur ;
- rotation de la tête du conducteur sur les côtés.

On constate que tous ces indices peuvent être extraits grâce aux algorithmes développés dans le cadre de la thèse d'Alexandre Benoit.

La détection de l'un ou plusieurs de ces indices conduit à l'envoi de messages d'alarmes au conducteur (voir Figure 57).

Ce workshop était particulièrement intéressant de part l'occasion qui nous a été offerte de travailler un mois durant avec des chercheurs d'origine différentes mais aussi car il nous a été possible de développer un système complet intégrant un certain nombre des algorithmes développés par ailleurs hors d'un cadre très précis d'application. Enfin, ce démonstrateur représente la base d'un système qui va continuer à être enrichi en insistant plus particulièrement sur l'analyse des signaux biologiques.

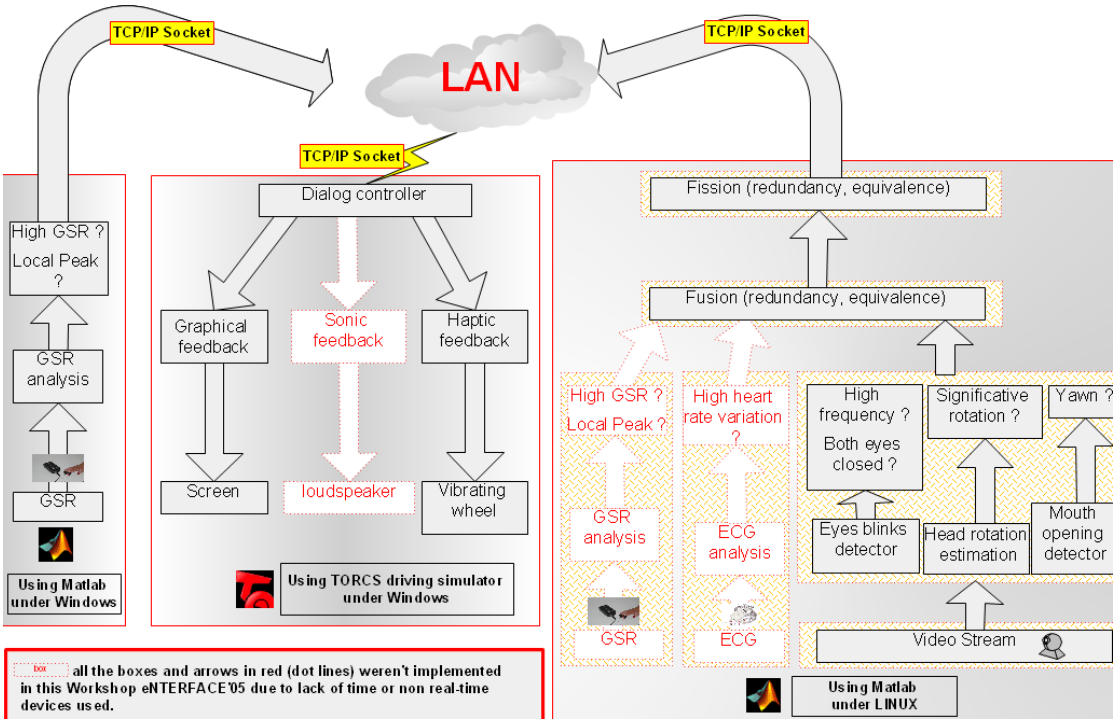

Figure 56 : architecture globale

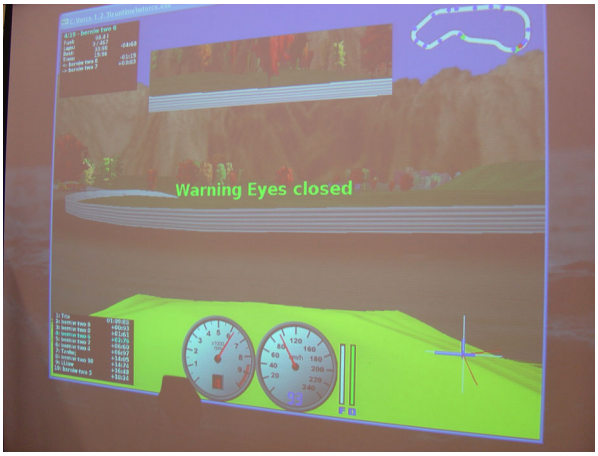

Figure 57 : exemple de message d'alerte visuel

9.5 Le projet DEIXIS (groupement GIS PEGASUS)

Le GIS PEGASUS (http://www.icp.inpg.fr/PEGASUS/) est un groupement d'intérêt scientifique (GIS) créé par convention le 15 décembre 2003 entre le CNRS et les 4 universités grenobloises (Institut National Polytechnique, Université Joseph Fourrier, Université Pierre Mendès France et Université Stendhal).
Le GIS PEGASUS est un réseau de recherche thématique centré sur les « objets, agents et environnements communicants ». Il vise à mettre en réseau les compétences des laboratoires de structures fédératives grenobloises (ELESA, IMAG, MSH Alpes) travaillant sur l'observation, la compréhension et la modélisation de comportements humains dans une situation d'interaction avec un système d'information central ou distribué médiatisée par des objets, agents ou environnements communicants.
Dans le cadre du groupement PEGASUS, le LIS est impliqué dans un projet de DEIXIS MULTIMODALE. Il s'agit de mettre en œuvre et valider une plate-forme de réalité partagée permettant non seulement de déterminer où se porte l'attention d'un usager dans le monde

physique mais aussi d'attirer son attention sur un endroit/objet précis… ceci médiatisé de manière efficace et intuitive par un agent conversationnel virtuel et à l'aide de dispositifs de poursuite non-invasifs (cf. Figure 58). Ce système devra permettre de répondre aux questions typiques d'un enfant de 12 mois à son tuteur adulte, en situation d'exploration de son univers immédiat : (a) où se trouve tel objet ? (b) quel est cet objet ?

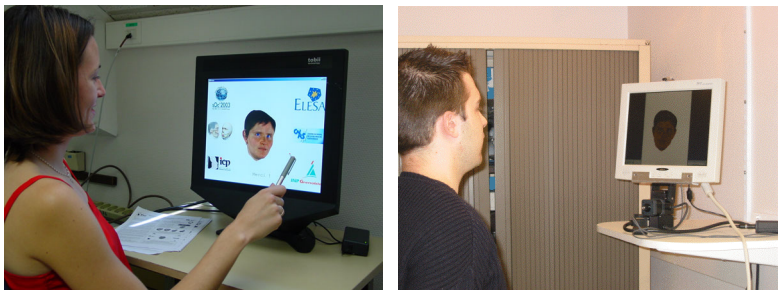

Figure 58 : plate forme MICAL de l'ICP : conversation entre un clone et un usager avec désignation d'objets

Dans le cadre de ce projet, le LIS intervient :

- au niveau de la détection et de la localisation de l'usager (cf. travaux de la thèse de Vincent Girondel sur la détection et le suivi de personnes en mouvement) ;
- au niveau de l'analyse de l'orientation de la tête et du bras de l'usager (cf. travaux de la thèse d'Alexandre Benoit et du DEA de N.Cacciaguera) ;
- au niveau de l'analyse de la direction du regard (cf. travaux de la thèse de Zakia Hammal et du DEA de Y. Lahaye)

9.6 Cluster régional « Informatique, signal et logiciels embarqués » : projet PRESENCE

Au niveau de la région Rhône-Alpes je participe au projet PRESENCE (Objets et Environnements Communicants, Interaction Homme-Machine et Usages) dont l'objectif est de rendre perceptifs, interactifs et communicants les objets, les agents artificiels et l'environnement.

Le projet PRESENCE vise à structurer les recherches autour des trois principaux enjeux socio-cognitifs des nouveaux dispositifs d'interaction.

- Le niveau « bas » de l'interaction sensorielle allant de la caractérisation psychophysique des dispositifs de perception et d'action jusqu'aux études comportementales et cognitives sur la boucle perception/action multimodale intégrant des dispositifs artéfactuels, ceci sans solliciter nécessairement un niveau d'interprétation sémantique complexe des signaux.
- Le niveau « intermédiaire » dans lequel les systèmes sollicitent les dimensions cognitives intégrant les dimensions linguistiques, affectives et sociales dans l'interprétation du comportement des usagers et la génération de leurs propres actions.
- Le niveau « haut » dans lequel les systèmes encore plus globaux visent à accompagner l'usager dans son activité quotidienne à la manière d'un ange-gardien phagocytant les objets communicants de l'environnement ou accompagnant l'usager dans ses déplacements lui permettant d'interagir de manière cohérente et située.

Le projet PRESENCE met particulièrement l'accent sur les plates-formes expérimentales permettant de réaliser et valider des scénarios d'usage autour de plateaux techniques instrumentés.

9.7 Développement d'une plate-forme interaction multimodale au LIS

L'intérêt expérimental et applicatif des recherches développées autour de l'analyse du mouvement humain en général ne peut être validé que si on met en œuvre dans un cadre de fonctionnement réaliste et temps réel les algorithmes développés. D'où l'idée de développer une plate-forme vidéo au laboratoire. Depuis 2003, nous travaillons au développement de cette plate-forme en définissant dans un premier temps les équipements nécessaires et en mettant en œuvre dans un second temps les algorithmes d'analyse et d'interprétation de mouvement humain ayant été éprouvés par ailleurs. Les bénéfices et les retours sur la recherche d'une telle plate-forme sont multiples :

- elle permet de tester plus aisément l'influence des paramètres associés aux différents algorithmes développés avec un retour immédiat de leur effet ;
- elle permet de prendre en compte la diversité des conditions d'éclairage ainsi que la diversité des comportements différents. En effet, lorsque l'on demande à un « cobaye » de tester une application de détection de hochements de tête de négation ou d'approbation par exemple, sa réalisation ne sera pas biaisée par les éventuels défauts de l'algorithme. Par ailleurs, ce genre d'expérience nous permet de prendre en compte des cas auxquels nous n'avions pas pensé. Ceci est très enrichissant car dans l'objectif du développement d'interfaces homme/machine, il est souhaitable de limiter les contraintes d'utilisation au risque sinon d'agacer l'utilisateur.
- elle permet de garder en mémoire l'importance de la contrainte d'un fonctionnement en temps réel (15 images par seconde au minimum sont nécessaires).

L'existence de cette plate-forme nous apparaît fondamentale pour nos recherches. Jusqu'ici trois stagiaires ingénieur en projet de fin d'études nous ont aidé pour les développements. En décembre 2005, il est prévu le recrutement d'un nouvel ingénieur de recherche dont le travail consistera entre autre au développement de cette plate-forme.

9.7.1 Description du matériel

La plate-forme est équipée du matériel suivant :

- 2 caméras SONY, DFW-VL 500 (cf. Figure 59) qui présentent l'avantage de permettre l'acquisition d'images à la cadence de 30 img/s en mode progressif non compressé. Les images peuvent être acquises en mode couleur ou niveau de gris et avoir pour taille 640x480 ou 320x240. L'ensemble des paramètres des caméras (zoom, focus, balance des blancs...) peut être réglé soit en mode automatique soit en mode manuel.
- une caméra stéréovision Bumblebee (cf. Figure 59)
- 2 tourelles (cf. Figure 59) pour asservissement en *pan* et *tilt* d'une caméra ou d'un écran
- 2 machines de traitement bi-processeurs à 3.2 GHz pour traitement temps-réel
- 1 micro cravate sans fil pour l'acquisition simultanée du son
- 4 projecteurs halogènes produisant un éclairage diffus (si besoin est)
- un vidéo-projecteur utilisé dans le cadre d'applications de réalité mixte
- un écran de projection

Figure 59 : de gauche à droite : caméra numérique, caméra stéréovision, tourelle

9.7.2 Exemples de démos temps réel

Cette plate-forme est utilisée pour le développement d'applications particulières. On distingue :

- le développement d'une plate-forme de réalité partagée permettant de savoir où se porte l'attention d'un sujet mais aussi d'attirer son attention sur un endroit ou un objet précis et de manière multi-modale.
- le développement d'une application d'analyse de mouvement multi-échelle (cf. Figure 60) : une première caméra filme en champ large de manière à permettre la détection et le suivi des personnes présentes dans la scène ; la seconde caméra asservie visuellement par la première se focalise sur le visage de la personne d'intérêt.
- le développement d'un système d'interprétation des gestes de la tête (hochement de tête, clignement d'yeux…)
- création d'une « smart room » ou pièce à ambiance intelligente : deux zones sont envisagées : une zone dite de travail dans laquelle l'utilisateur fait face à un PC et peut interagir avec ce PC de manière intuitive et une zone dite de loisir dans laquelle l'utilisateur est sur un canapé et il interagit avec l'environnement par des gestes, par la voix … afin par exemple de projeter un film sur l'écran par l'intermédiaire du vidéo projecteur.

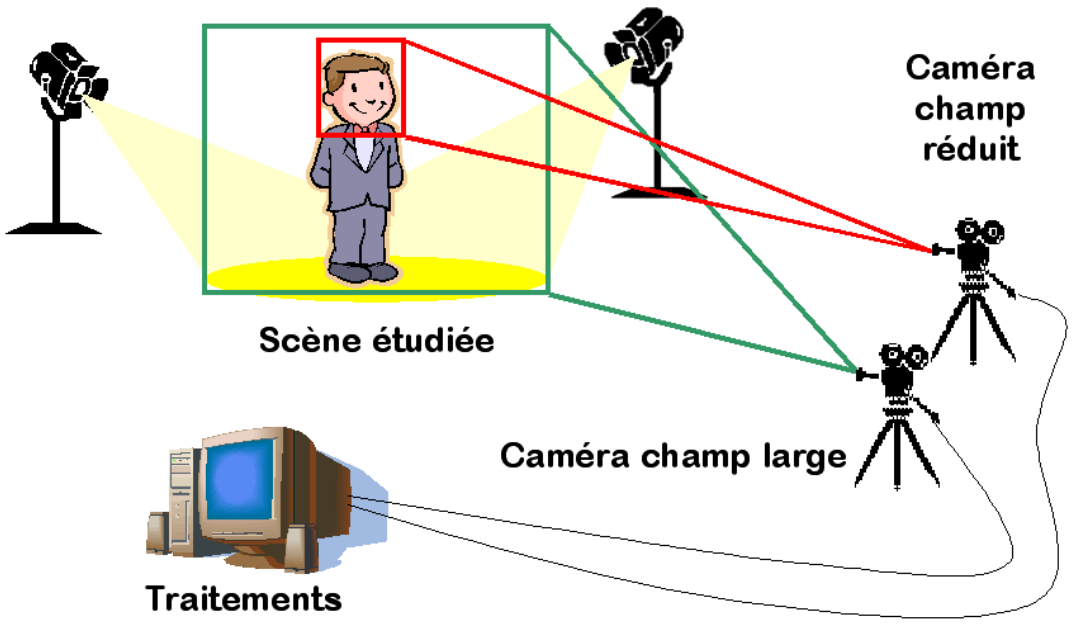

Figure 60 : analyse muti-échelle du comportement d'un utilisateur

10 Projet de recherche : analyse et interprétation multi-modales d'activités humaines

Pour la suite de mes recherches, je propose de développer une activité autour de la thématique des activités humaines assistées par ordinateur. L'utilisation des moyens informatiques a envahi notre quotidien. Citons pour exemple le téléphone portable, l'assistance à la conduite ou à l'orientation dans les voitures, les systèmes de reconnaissance par biométrie (citons à titre anecdotique le cas de cette machine à laver proposée en Espagne qui identifie les empreintes de son utilisateur avant de permettre l'accès au chargement). Donc nous sommes de plus en plus amenés à être informés, à communiquer ou à interagir avec des systèmes informatiques avec tout ce que cela comporte comme difficultés ou peur de la part de tout un chacun. Afin de limiter ces difficultés, il serait bénéfique d'intégrer les spécificités et les caractéristiques des « comportements » humains ainsi que de leur mode de communication.
Par rapport au domaine de recherche envisagé, on peut distinguer deux grands aspects :

- l'analyse et l'interprétation des « activités » humaines sur la base d'un ensemble de signaux émanant de l'humain (parole, expressions, gestes, comportement…) ;
- l'interaction entre l'homme et la machine qui découle de l'étape précédente.

Dans le cadre du projet de recherche proposé, nous nous intéressons avant tout au premier aspect.
Indiquons en remarque préalable que le Laboratoire des Images et des Signaux où j'ai mené mes recherches jusqu'à présent est en passe de regroupement avec deux autres laboratoires de Grenoble à savoir l'institut de la Communication Parlée et le Laboratoire d'Automatique de Grenoble. Nous verrons que le projet de recherche proposé s'intègre tout à fait dans cette perspective car il s'appuie en partie sur des compétences supplémentaires qui émergeront de cette fusion.

10.1 Analyse et interprétation multi-modales des activités humaines sur la base de signaux multi-capteurs.

Le diagramme de la Figure 61 présente une vue d'ensemble de l'organisation du projet proposé.

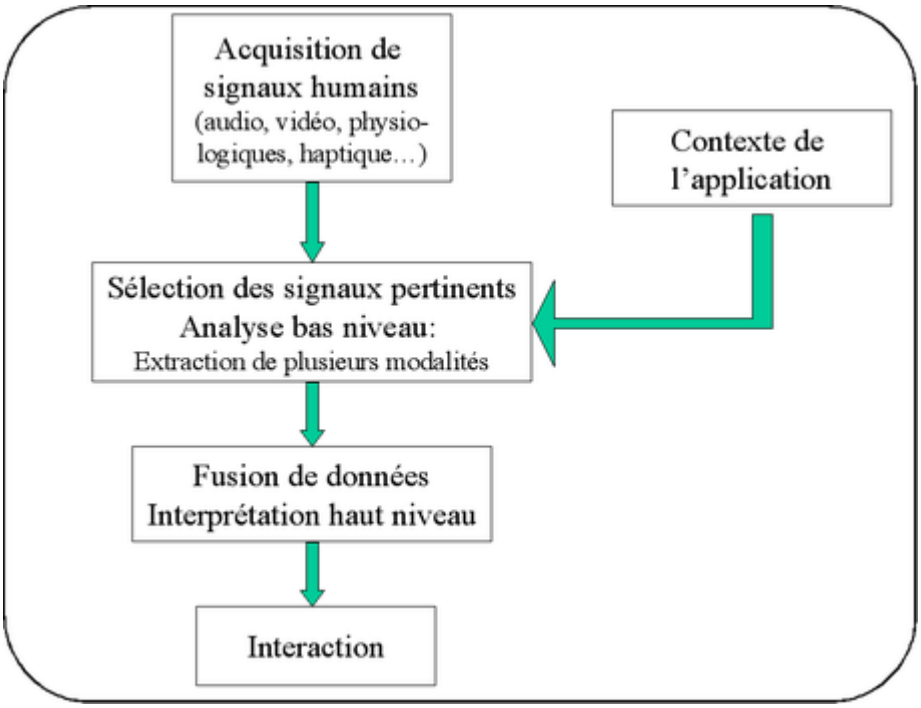

Figure 61 : vue d'ensemble du projet proposé

10.1.1 Capteurs

Jusqu'ici les seuls capteurs ayant été utilisés sont des caméras, soit des caméras monoculaires soit des caméras stéréo. On envisage d'enrichir la source des signaux disponibles par :

- l'acquisition de l'audio : il est évident que de nombreuses informations passent par le canal audio. On s'intéressera non seulement au message de parole (reconnaissance de mots et de phrases) mais aussi aux informations véhiculées par les intonations (une analyse des expressions dans le signal de parole a déjà été initiée). Pour ce faire, nous appuierons bien évidemment sur les compétences du grand laboratoire en cours de création (cf. équipes « machines parlantes » et « structure du code » de l'actuel Institut de la Communication Parlée).
- l'acquisition de signaux physiologiques : des informations liées à l'activité humaine sont véhiculées par un certain nombre de signaux physiologiques tels que l'EEG, la GSR, l'EMG, la fréquence cardiaque, la fréquence respiratoire. L'idée est alors d'acquérir une centrale d'acquisition de tels signaux. Le gros avantage mais aussi le principale difficulté des signaux physiologiques par rapport au signal vidéo ou audio résident dans le fait qu'il n'est pas possible a priori d'en contrôler consciemment les variations. Plus précisément, autant il est possible de simuler sur une vidéo toutes les caractéristiques d'un visage en colère ou bien d'un endormissement, autant il n'est pas possible de simuler l'EEG d'une personne endormie.

Pour l'analyse des signaux physiologiques, nous nous appuierons sur les compétences du laboratoire en cours de création à savoir l'équipe « Sûreté et diagnostique » du Laboratoire d'Automatique de Grenoble.

Remarquons que l'acquisition de l'audio et de la vidéo nécessitent des systèmes a priori moins encombrants et moins intrusifs que les autres signaux considérés.

10.1.2 Sélection des signaux pertinents et analyse bas-niveau

La seconde étape, la plus délicate à mon avis, concerne l'analyse des signaux acquis afin d'en extraire des informations de bas niveau que l'on appellera modalités dans ce qui suit. Avant cette analyse, on se posera la question de savoir si tous les signaux acquis sont pertinents. Cette sélection est tributaire de l'application considérée et elle doit pouvoir se faire soit de manière a priori, soit en cours d'utilisation (et ce, soit de manière automatique, soit par intervention de l'utilisateur). Par exemple, si le système envisagé est destiné à des personnes mal-entendantes profondes comme cela peut être le cas pour le terminal du projet Telma, l'acquisition des signaux de parole issus de l'utilisateur n'a pas de sens.
Ensuite, l'analyse des signaux issus de chaque type de capteur fournit un ensemble de **modalités**. La liste ci-dessous fournit des exemples de modalités possibles issues de l'analyse des signaux de chacun des capteurs envisagés :

- capteur vidéo : mouvement, contours et déformations de traits faciaux, position et mouvement de la tête, position des mains, nombre et localisation de chaque acteur…
- capteur audio : message, énergie du signal (niveau sonore), pitch (proportions des hautes et des basses fréquences), contenu fréquentiel, débit de parole…
- capteur physiologique : analyse statistique des variations du rythme cardiaque, des variations de la conductivité de la peau, type d'ondes présentes dans les EEG…

La phase d'analyse consiste à extraire des signaux disponibles le maximum d'informations. Chaque information extraite est en générale une information dite de bas niveau au sens où il n'est pas possible avec chaque information prise séparément d'en déduire une interprétation sémantique sur les « activités » de l'utilisateur.
L'étape d'analyse des signaux acquis est une étape clef pour la suite du système. Elle requiert des algorithmes robustes et fonctionnant en temps réel.

10.1.3 Fusion et interprétation haut niveau

Une fois les modalités sélectionnées et extraites, il s'agit de les fusionner afin d'aboutir à une interprétation de haut niveau à savoir à une interprétation sémantique. Par exemple, pour détecter un état d'hypovigilance chez un conducteur automobile, on fusionne les modalités ouverture des yeux, mouvements de la tête, direction du regard, détection de bâillements répétés. L'étape de fusion pourra éventuellement être vue comme un processus de classification. Ceci nous amènera à poursuivre et étendre notre collaboration avec les chercheurs du LIS travaillant dans le domaine de la classification et de la fusion de données afin d'apporter des solutions aux problèmes suivants :

- toutes les modalités extraites dans la phase d'analyse bas niveau sont elles compatibles ? A savoir est il possible de les considérer toutes dans un même processus de fusion sachant qu'elles peuvent provenir de sources différentes ? A titre d'illustration de ces propos, si on considère le cas de la reconnaissance des émotions chez un utilisateur, on a vu que les modalités issues de la vidéo conduisent à une classification en 7 classes (six émotions universelles + émotion neutre) alors que les modalités issues de l'audio nous ont conduit à une classification en deux classes seulement (active et passive). On constate donc que la fusion des modalités audio et vidéo n'est a priori pas immédiate : toutes les modalités ne pourront pas être traitées de la même manière. Il s'agira de définir une stratégie de fusion adaptée.
- toutes les modalités extraites doivent elles être traitées de la même manière ? Par exemple, si on s'intéresse à l'état de vigilance d'un conducteur, la détection de la modalité « yeux fermés » devrait être prépondérante sur beaucoup d'autres. En effet, il est clair que si le conducteur a les yeux fermés, il y a danger immédiat. D'où l'idée de pouvoir, dans le processus de fusion, donner plus de poids à une modalité par rapport à une autre.

- existe il des modalités redondantes ? Si oui, est il nécessaire de les considérer toutes ou alors faut il envisager des voies de fusion parallèles qui permettraient de confirmer ou d'infirmer une interprétation ?
- la fusion suppose implicitement un apprentissage ou une expertise afin de caractériser en terme de modalités (et fusion de modalités) les « actions humaines » à reconnaître. Par exemple, le processus de fusion de données présenté au paragraphe 5.2 pour la reconnaissance des émotions s'appuient sur le Tableau 1 qui définit les évolutions attendues de chacune des modalités considérées en relation avec les émotions recherchées. Cette phase d'apprentissage ou d'expertise peut être relativement complexe et elle nécessite éventuellement des études psychologiques préalables (comme par exemple, celle que nous avons menée pour valider l'hypothèse selon laquelle les expressions faciales sont reconnaissables à partir de la visualisation des déformations des traits du visage uniquement). En particulier, il nous faudra acquérir une expertise sur l'évolution des signaux physiologiques par rapport à une activité humaine donnée.
- faut il faire une fusion au niveau de l'ensemble des modalités (« *feature level fusion* ») ou alors doit on envisager une fusion au niveau d'un ensemble de décisions intermédiaires (« *decision level fusion* ») ? Dans le cas évoqué de la reconnaissance des émotions, on pourrait soit envisager une fusion de l'ensemble des informations issues de l'audio et de la vidéo afin d'aboutir à la classification ou alors faire une classification à partir des informations audio, une classification à partir des informations vidéo puis fusionner ensuite chacune des décisions intermédiaires pour aboutir à la décision finale.
- l'ensemble des modalités étant issues non pas directement de l'information donnée par le capteur mais d'une analyse préalable de ces informations, il existera nécessairement des modalités qui seront erronées. Il est donc nécessaire d'envisager des procédés de fusion de données capables de tenir compte de données erronées ou incomplètes.
- toutes les interprétations envisagées se situent dans un cadre dynamique si bien que la prise en compte de la dimension temporelle dans le processus de fusion de données est indispensable. Ceci soulèvera des problèmes théoriques tels que par exemple, comment étendre la théorie de l'évidence à un cadre dynamique. Comment également tenir compte des problèmes de synchronisation ?

10.1.4 Interaction

Le but de l'ensemble de la chaîne réside dans cette étape d'interaction avec le système ou d'adaptation du système à l'utilisateur. Par exemple, dans le cas d'un serveur vocal, on pourrait envisager que le débit du serveur diminue dès lors qu'il a détecté que l'interlocuteur est une personne âgée ou que l'intonation de la voix de synthèse soit apaisante dans le cas d'un utilisateur énervé.

Les actions ou interactions que nous pourrons proposer dans le cadre de ce projet de recherche seront très basiques. Nous ne prétendons pas développer des recherches poussées dans le domaine de l'interaction homme machine. Si besoin est, nous nous appuierons sur les travaux de certains partenaires du réseau d'excellence Similar tels que le laboratoire CLIPS (équipe IHM) de Grenoble par exemple.

10.2 Analyse des activités humaines : exemples d'applications

Afin d'illustrer la démarche exposée dans la section 10.1, nous proposons un ensemble d'applications. Parmi les exemples proposés, certains sont déjà en cours de développement au laboratoire et d'autres non.

10.2.1 Immersion dans un environnement virtuel

A l'instar de ce que nous avons ébauché lors du projet Art-live, il s'agit de développer des systèmes pour lesquels l'utilisateur est extrait de l'environnement réel pour être immergé dans un environnement virtuel dans lequel il se voit évoluer et dans lequel il peut interagir par gestes, par la parole... Les capteurs envisagés pour ce type d'application sont la vidéo, l'audio et les systèmes haptiques. La démarche consiste alors à définir a priori un ou plusieurs scénarios d'interaction ce qui se traduira par la définition d'un ensemble de caractéristiques à détecter lors de l'évolution de l'utilisateur. Il faudra alors définir les modalités à fusionner pour pouvoir détecter automatiquement ces caractéristiques ainsi que les capteurs à utiliser. Par exemple, lors du projet Art-live, il était demandé à l'utilisateur d'attraper un ensemble de papillons virtuels ce qui était fait lorsque la main de l'utilisateur se trouvait sur un papillon. Pour ce faire, le capteur utilisé était une caméra monoculaire dont on a extrait les modalités de position globale de la personne et de position de ses mains à partir de l'analyse des images fournies par la caméra.
On peut aussi bien imaginer que l'utilisateur soit immergé dans un environnement virtuel sonore dont le contenu évoluerait en fonction d'une part de son activité mais aussi de son état d'esprit (idée d'une ambiance sonore contrôlée par les émotions).

10.2.2 Surveillance de l'état d'attention ou de stress d'un utilisateur

De nombreuses tâches nécessitent l'attention de l'utilisateur soit pour des raisons de sécurité soit pour des raisons d'efficacité. Citons par exemple le cas d'un conducteur automobile qui se doit d'être attentif à la route et le cas d'un élève en situation d'apprentissage à distance qui se doit d'être concentré s'il veut être efficace dans son apprentissage. Pour la détection automatique des états d'inattention ou de stress de ce type d'utilisateur, on peut envisager la fusion de modalités issues de la vidéo (bâillement, fermeture des yeux, hochements de tête) et de signaux physiologiques (EEG, battement du coeur, fréquence respiratoire). La détection automatique de tels états permettrait une « intervention » soit du système pour réveiller le conducteur (cf. projet Enterface) soit d'une personne telle le professeur pour apporter de l'aide à l'élève inattentif ou en difficulté.

10.2.3 Systèmes pour handicapés : compensation du handicap.

Dans le cas de certains handicaps, certains signaux humains sont déficients. Par exemple, un aveugle ne dispose pas de l'image, un sourd ne dispose pas du son. L'idée est donc de développer des systèmes qui permettraient de remplacer, de compléter ou de suppléer les informations manquantes et de pouvoir par l'intermédiaire de l'outil informatique remplacer une modalité sensorielle par une autre. Il est probable que le remplacement de l'information déficiente nécessite la fusion de plusieurs modalités. Le terminal Telma en est un bon exemple. De fait, une personne malentendante n'a pas accès au téléphone. En revanche, on a montré que ceci pouvait être théoriquement possible en ajoutant des modalités issues du canal vidéo, le signal de parole reçu ou émis pouvant être traduit et transformé en message vocal. Les modalités nécessaires pour ce faire sont multiples : configuration et position de la main pour l'identification de la syllabe codée via le LPC, contours de la bouche et paramètres de formes associés.

10.2.4 Détection de comportements à risque

Un autre type d'application serait la détection de comportements particuliers d'un ou de plusieurs personnes tels que les situations d'agression dans les transports en commun par exemple. La détection de ce genre de situation s'inscrit tout à fait dans notre démarche globale. Il faut de nouveau définir les capteurs et les modalités pertinentes (ici l'audio, la

vidéo, la fréquence cardiaque du chauffeur) et caractériser les comportements recherchés vis à vis des modalités sélectionnées.

10.3 Conclusion

Ce que nous proposons avant tout dans le cadre de ce projet de recherche consiste en la caractérisation des activités humaines au moyen de l'analyse de signaux « humains » issus de différents capteurs (audio, vidéo, physiologique). Notre travail portera d'une part sur l'étape d'analyse des signaux qui est une étape clef et qui nécessite des algorithmes robustes et fonctionnant en temps réel et d'autre part sur les processus de fusion de données.
Le projet proposé s'inscrit dans la politique scientifique du LIS. Il s'inscrit aussi dans le cadre d'une collaboration plus forte avec certaines équipes du LAG et de l'ICP en prévision de la fusion en 2007 de nos trois laboratoires ce qui est illustré par le schéma de la Figure 62.

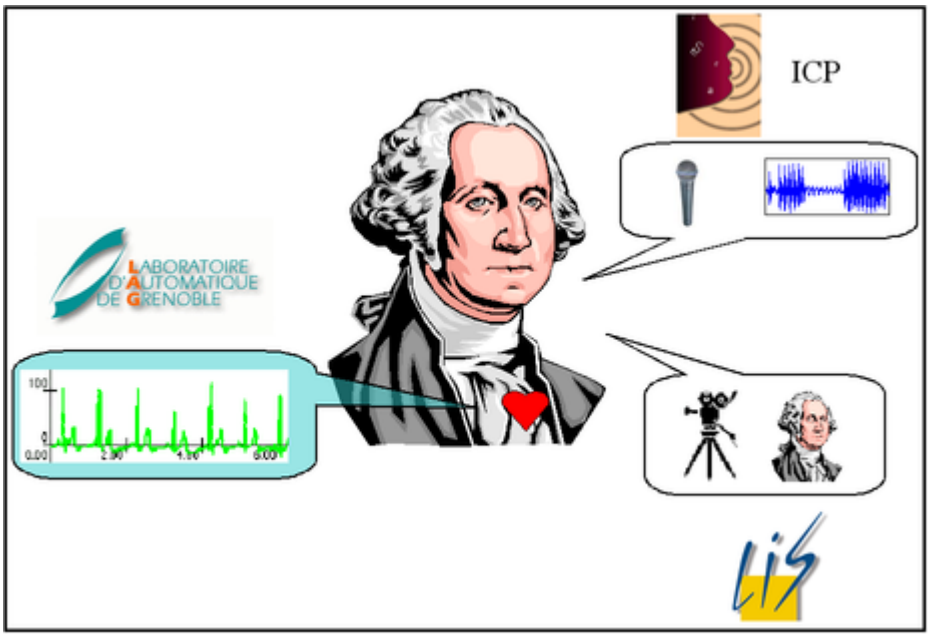

Figure 62 : analyse de signaux liés à l'humain dans le contexte du regroupement LIS, LAG, ICP.

Sur le plan national, le projet de recherche proposé s'inscrit directement dans les nouvelles thématiques du GDR-ISIS avec l'apparition d'un axe « visage et gestes » lors de la dernière assemblée générale du groupement.
Sur le plan international, le projet de recherche proposé s'inscrit dans le cadre de la thématique générale sur les interactions multi-modales développée par le réseau d'excellence Similar dont le LIS est un partenaire actif.

Références

[Abboud04] B. Abboud, F. Davoine, Facial expression recognition and synthesis based on an appearance model, *Signal Processing: Image Communication*, Elsevier, Vol. 19, No. 8, pages 723-740, September, 2004.

[Alleysson99] D. Alleysson – Le traitement du signal chromatique dans la rétine. Un modèle de base pour la perception humaine des couleurs- *Thèse de l'UJF, spécialité : informatique, sciences cognitives*, laboratoire LIS, 1999.

[Aggarwal03] J.K. Aggarwal, Q. Cai – Human motion analysis : a review – *Computer Vison and Image Understanding*, vol.73, n°3, pp.428-440, 1999.

[Bassili78] J.N. Bassili. - Facial motion in the perception of faces and of emotional expression- *Journal of Experimental Psychology, vol.4, pages 373-379, 1978.*

[Beaudot94] W. Beaudot -The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision- *PhD Thesis in Computer Science*, INPG (France) december 1994

[Bobick01] A. Bobick, J. Davis –The Recognition of Human Movement Using Temporal Templates- IEEE Trans. on PAMI, Vol.23, n°3, pp.257-267, mars 2001.

[Buciu03] I. Buciu, C. Kotropoulos, I. Pitas. ICA and Gabor representation for facial expression recognition. ICIP (2) 2003: 855-858

[Chesnaud00] C; Chesnaud – Techniques statistiques de segmentation par contours actifs et mise en oeuvre rapide- *Institut Fresnel, thèse de l'université Aix-Marseille*, 2000.

[Cohn98] J. Cohn, A.J. Zlochower, J.J. Lien , T. Kanade, -Feature-point tracking by optical flow discriminates subtles differences in facial expression- *IEEE International Conference on automatic Face and Gesture Recognition, N°3, pages 396-401, 1998.*

[Cohn] Facial expression database http://www-2.cs.cmu.edu/~face

[Collins00] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y; Tsin, D. Tolliver, N. Enomoto, O. Hasegawa – A system for video surveillance and monitoring – *CMU-RI-TR*, 2000.

[Cootes95] T.F Cootes, C.J. Taylor, D. Cooper and J. Graham - Active Shape Models: their Training and Application - *In Computer Vision and Image Understanding*, 1(61), pp.38-59, January 1995.

[Dailey01] M. Dailey, G.W. Cottrell, & J. Reilly, - California Facial Expressions - *CAFE, unpublished digital images, UCSD Computer Science and Engineering Department, 2001*

[Dempster68] A. Dempster. - A generalization of Bayesian inference - *Journal of the royal statistical society, vol.30, pages 205-245, 1968.*

[DES] Audio emotions data base http://cpk.auc.dk/~tb/speech/Emotions/

[Edwards98] G.J. Edwards, T.F Cootes, C.J. Taylor – Face recognition using active appearance models – Proc. European Conf. Computer Vision, vol.2, pp. 581-595, 1998.

[Ekman78] P. Ekman, W. Friesen – Facial Action Coding System (FACS): Manual- *Palo Alto: Consulting Psychologist Press*, 1978.

[Ekman99] P. Ekman, -Facial Expression-, *The Handbook of Cognition and Emotion, T. Dalgeleish and M. Power, John Wiley & Sons Ltd. 1999.*

[Fasel03] B. Fasel, J. Luettin – Automatic Facial Expression analysis: a survey- *Pattern Recognition*, vol.36, pp.259-275, 2003.

[Forster01] F. Forster, M.Lang, B.Radic -Real-time 3D and color camera- *In Proc. ICAV3D 2001*, Mykonos, Greece, May 2001

[Gravilla99] D. Gravilla – The Visual Analysis of Human Movement : a Survey- Computer Vision and Image Understanding, vol.73, n°1, pp. 82-98, 1999.

[Haritoaglu98] I. Haritoaglu, D. Harwood, L. Davis – Who, when, where, what: a real time system for detecting and tracking people- *IEEE International conference on Automatic Face and Gesture Recognition*, April, pp.222-227, 1998.

[Hjelmäs01] H. Hjelmäs, B. Low - Face detection: a survey - *Computer Vision and Image Understanding*, 83, pp. 236-274, 2001.
[Justin03] P. Juslin, P. Laukka -Communication of emotions in vocal expression and music performance: Different channels, same code?- *Psychological Bulletin*, pp. 770-814, 2003.
[Malciu01] M. Malciu - Approche orientées modèle pour la capture des mouvements du visage en vision par ordinateur - *Thèse de l'université René Descartes, Paris V, Decembre 2001.*
[Mehrabian68] A. Mehrabian, - Communication without words - *Psychology , vol 4, pp. 53-56, 1968.*
[MPT] Machine Perception Toolbox, face detection algorithm:
http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/introductionframe.html
[Oliva99] A. Oliva, A. Torralba, A. Guérin, J. Hérault. -Super-Ordinate representation of scenes from power spectrum shapes-, *CIR-99, The challenge of image retrieval*, Newcastle, March 1999.
[Ong05] S. Ong, S. Ranganath – Automatic Sign Language Analysis: A survey and the Future beyond Lexical Meaning – IEEE trans. on PAMI, vol.27, n°6, pp. 873-891, June 2005.
[OpenCV] Opencv website : http://sourceforge.net/projects/opencvlibrary
[Pantic00a] M. Pantic, L.J.M Rothkrantz –Automatic analysis of facial expressions: the state of the art- *IEEE trans. on PAMI*, vol.22, N°12, pp. 1424-1445, December 2000.
[Pantic00b] M. Pantic, L.J.M Rothkrantz - Expert system for automatic analysis of facial expressions - *ElSEVIER Image and Vision Computing, vol.18, pages 881-905, 2000.*
[Pavlovic97] V. Pavlovic, R. Sharma, T. Huang – Visual interpretation of hand gestures for Human computer Interaction: A review – *IEEE trans. on PAMI*, vol.19, n°7, pp. 677-695, July 1997.
[Ritcher82] J. Ritcher, S.Ullman. -A model for temporal organization of X- and Y-type receptive fields in the primate retina.- *Biological Cybernetics*, 43:127-145,1982.
[Scherer03] K. Scherer -Vocal communication of emotion- *A review of research paradigms. Speech Communication,* Vol 40, pp. 227-256, 2003.
[Schröder03] M. Schröder, -Speech and Emotion Research-, *PHD thesis report*, university of Saarlandes, 2003.
[Shafer76] G. Shafer. - A Mathematical Theory of Evidence - *Princeton University Press, 1976.*
[Similar] Similar Web Site : www.similar.cc
[Smets98] P. Smets - Handbook of Defeasible Reasoning and Uncertainty Management Systems: The transferable belief model for quantified belief representation - v*ol. 1, pages 267-301, publisher Kluwer, Doordrecht, The Netherlands 1998*
[Sourd03] A. Sourd - Etude du processus de reconnaissance émotionelle à partir de squelettes faciaux d'émotions - *Mémoire de maîtrise de psychologie, laboratoire LPS, 62 pages, 2003.*
[Terrillon00] J.C. Terrillon, M. Shirazi, H. Fukamachi, S. Akamatsu – Comparative performances of different skin chrominance models and chrominance spaces for the automatic detection of human face sin color images- *Inter. Conf . on Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 54-61.
[Torralba99] A. Torralba –Architectures analogiques pour le traitement d'images : réseaux cellulaires neuronaux et circuits neuromorphiques- *Thèse de l'INPG, spécialité Signal, Image, Parole*, laboratoire LIS, 1999.
[Viola04] P. Viola, J. Jones - Robust Real Time Face Detection- *International Journal of Computer Vision*, 57(2):137-154, May, 2004.
[WangJ03] J.J. Wang, S. Singh – Video analysis of human dynamics: a survey – *Real Time Imaging*, vol.9, pp.321-346, 2003.
[WangL03] L. Wang, W.Tan – Recent developments in human motion analysis – *Pattern Recognition*, vol.36, n°3, pp.585-601, 2003.
[Wreng97] C.R. Wreng, A. Azarbayejani, T; Darell, A. Pentland – Pfinder: real time tracking of the human body – *IEEE trans. on PAMI*, vol.19(7), pp. 780-795, July 1997.
[Yang02] M.H. Yang, D. Kriegman, and N. Ahuja - Detecting face in images : a survey - *IEEE Trans on PAMI*, vol.24, n°1, pp. 34-58, January, 2002

Publications

Rque : toutes les publications ont été classées par ordre chronologique. Trois classes ont été définies :

- classe [R_nomdate] pour les articles de revue à comité de lecture
- classe [CI_nomdate] pour les articles de conférences internationales
- classe [CN_nomdate] pour les articles de conférences nationales

Revues avec comité de lecture

[R_Hammal05a] Z. Hammal, N. Eveno, A. Caplier., P.Y. Coulon - Parametric models for facial features segmentation- A paraître dans *Eurasip Journal of Signal Processing en 2005.*

[R_Hammal05b] Z.Hammal, N. Eveno, A. Caplier, P.Y. Coulon - Extraction des traits caractéristiques du visage à l'aide de modèles paramétriques adaptés – *Traitement Du Signal*, volume 22, N°1, pp.59-72, 2005.

[R_Girondel05] V. Girondel, A. Caplier, L. Bonnaud ,M. Rombaut, -Belief theory-based classifiers comparison for static human body postures recognition in video- *International Journal of Signal Processing* IJSP, Vol. 2, n°1, pp-29--33, 2005.

[R_Eveno04] N. Eveno, A. Caplier, P.Y. Coulon - Automatic and Accurate Lip Tracking. *IEEE Transactions on Circuits and Systems for Video technology,* Vol.14, N°.5, mai 2004, pp.706-715

[R_Girondel03] V. Girondel, L. Bonnaud, A.Caplier Hands detection and tracking for interactive multimedia applications. *Archives of Theoretical and Applied Informatics*, 2003.

[R_Luthon99] F. Luthon, A. Caplier, M. Lievin - Spatiotemporal MRF approach to video segmentation : application to motion detection and lips segmentation.- *Signal Processing*, 76, 1999, pp.61-80

[R_Caplier98] A. Caplier, F. Luthon, C. Dumontier -Real Time Implementations of an MRF-based Motion Detection Algorithm. *Journal of Real Time Imaging, Special Issue on Real-Time Motion Analysis*, 4(1), February 1998, pp.41-54.

[R_Luthon97] A. Caplier, F. Luthon - Approche spatio-temporelle pour l'analyse de séquences d'images. Application en détection de mouvement. *Traitement du signal*, vol.14, n°2, 1997, pp.195-208.

[R_Caplier96] A. Caplier, C. Dumontier, F. Luthon, P.Y. Coulon.- Algorithme de segmentation de mouvement par modélisation markovienne. Mise en oeuvre sur DSP-. *Traitement du signal*, Volume 13, n°2, 1996, pp. 177-190

Conférences internationales

[CI_Benoit05a] A. Benoit, A. Caplier –Motion Estimator Inspired from Biological Model for Head Motion Interpretation – *WIAMIS05*, Montreux, Suisse, avril 2005.

[CI_Benoit05b] A. Benoit, A. Caplier - Biological Approach for Head Detection and Analysis - *EUSIPCO2005*, Antalya, Turkey, September 2005.

[CI_Benoit05c] A. Benoit, A. Caplier - Motion Head Nods analaysis: interpretation of non verbal communication gestures - *ICIP2005*, Genova, Italie, Septembre 2005

[CI_Benoit05d] A. Benoit, A. Caplier - Hypovigilence Analysis: Open or Closed Eye or Mouth ? Blinking or Yawning Frequency ?- *IEEE AVSS, International conference on Advanced Video and Signal based Surveillance*, Como, Italy, September 2005

[CI_Burger05] T.Burger, A. Caplier, S. Mancini – Cued Speech hand gestures recognition tool - *EUSIPCO2005*, Antalya, Turkey, September 2005

[CI_Enterface05] A. Benoit, L. Bonnaud, A. Caplier, P. Ngo, L. Lawson, D. Treviesan, V. Levacic, C. Mancas, G. Chanel – Multimodal Focus Attention Detection in an Augmented Driver Simulator- *eNTERFACE'05 workshop*, Mons, Belgium, July 18th, August 12th, 2005

[CI_Girondel05a] V. Girondel, L. Bonnaud, A. Caplier, M. Rombaut, -Belief theory-based classifiers comparison for static human body postures recognition in video-, *Proceedings of The Second World Enformatika Congress WEC'05*, Volume 1, pp. 237--240, February 25-27, Istanbul, Turkey, 2005.

[CI_Girondel05b] V. Girondel, A. Caplier, L. Bonnaud ,M. Rombaut. - Real Static Human Body Posture Recognition in Video Sequence Using the Belief Theory. - *ICIP2005*, Genova, Italy, September, 2005

[CI_Girondel05c] V. Girondel, A. Caplier, L. Bonnaud – A belief theory-based static postures recognition system for real-time video surveillance applications - *IEEE AVSS, International conference on Advanced Video and Signal based Surveillance*, Como, Italy, September 2005

[CI_Hammal05a] Z. Hammal, B. Bozkurt, L. Couvreur, D. Unay, A. Caplier, T. Dutoit- Quiet versus agitated: vocal classification system – *13th EUSIPCO*, Antalya, Turquie, Septembre 2005

[CI_Hammal05b] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut –Facial expression recognition based on the belief theory : comparison with different classifiers.- 13th ICIAP, Cagliary, Italy, Septembre 2005

[CI_Hammal05c] Z. Hammal, A. Caplier, M. Rombaut – A fusion process based on belief theory for classification of facial basic emotions - *8th International conference on Information Fusion*, Philadelphia, USA, July 2005

[CI_Hammal05d] Z. Hammal, C. Massot, G. Bedoya, A. Caplier -Eyes segmentation applied to gaze direction and vigilance estimation- 3rd International Conference on Advances in Pattern Recognition, Bath, United Kingdom, August 2005

[CI_Hammal05e] Z. Hammal, A. Caplier, M. Rombaut . -Belief Theory Applied to Facial Expressions Classification- *3rd International Conference on Advances in Pattern Recognition*, Bath, United Kingdom, August 2005

[CI_Caplier04] A. Caplier, L. Bonnaud, S. Mallassiotis, M. Strintzis - Comparison of 2D and 3D Analysis For Automated Cued Speech Gesture Recognition – *SPECOM04*, St Petersbourg, Russie, Septembre 2004

[CI_Girondel04] V. Girondel, A. Caplier, L. Bonnaud - Real Time Tracking of Multiple Persons by Kalman Filtering and Face Pursuit for Multimedia Applications *IEEE Southwest Symposium on Image Analysis and Interpretation*, USA, March, 2004

[CI_Hammal04] Z. Hammal, A. Caplier - Eyes and eyebrows parametric models for automatic segmentation -*IEEE Southwest Symposium on Image Analysis and Interpretation*, USA, March, 2004

[CI_Harasse04] S. Harasse, L. Bonnaud, A. Caplier, M. Desvignes - Automated Camera Dysfunctions Detection - *IEEE Southwest Symposium on Image Analysis and Interpretation*, USA, March, 2004

[CI_Eveno03] N. Eveno, A. Caplier, P.Y. Coulon. -Jumping snakes and parametric model for lip segmentation-. *IEEE International Conference on Image Processing*, Barcelone, Espagne, Sept. 2003

[CI_Rebuffel03] V. Rebuffel, S. Pires, A. Caplier, P. Lamarque -Automatic delamination defect detection in radiographic sequences of rocket boosters - *International Symposium on Computed Tomography and Image Processing for Industrial Radiology*, Berlin, june 2003

[CI_Artlive02] Artlive Consortium –The artlive architecture for mixed reality- *Virtual Reality International Conference*, Laval, France, June 2002

[CI_Eveno02a] N. Eveno, A. Caplier, P.Y. Coulon. -A parametric model for realistic lip segmentation- *International Conference on Control, Automation, Robotics and Vision (ICARV'02)*, Singapore, December 2002

[CI_Eveno02b] N. Eveno, A. Caplier, P.Y. Coulon. -Key points based segmentation of lip.- *IEEE International Conference on Multimedia and Expo*, August 26-29, 2002, Lausanne, Switzerland

[CI_Girondel02] V. Girondel, L. Bonnaud, A.Caplier - Hands Detection and Tracking For Interactive Multimedia Applications- , *International Conference on Computer Vision and Graphics*, 25-29 September, 2002, Zakopane, Poland
[CI_Artlive01] Artlive Consortium –Immersive Interatives narratives- *Proc of ICAV3D*, May 30-June 01, Ornos, Mykonos, Greece, 2001
[CI_Caplier01a] A.Caplier , L. Bonnaud, J.M. Chassery - Robust fast extraction of video objects combining frame differences and adaptive reference image *IEEE ICIP2001, International Conference on Image Processing*, Greece, Oct. 2001
[CI_Caplier01b] A.Caplier -Lip detection and tracking.- *11th International Conference on Image Analysis and Processing*, Palermo, Italy , September 26-28, 2001
[CI_Eveno01] N. Eveno, A. Caplier, P.Y. Coulon -A new color transformation for lip segmentation- *In Proc. IEEE MSSP'01*, Cannes, France, September 2001
[CI_Caplier00] A. Caplier - Adaptation of ASM to lip edge detection.- *Articulated Motion and Deformable Objects ADMO'2000*, Palma de Majorque, Espagne, septembre 2000, pp.27-37.
[CI_Caplier95a] A. Caplier, F. Luthon -An MRF-based Spatiotemporal Multiresolution Algorithm for Motion Detection.- *Proc. SCIA95*, Upsalla, Suède, Juin 1995, pp.559-566
[CI_Caplier95b] A. Caplier, F. Luthon. -Spatiotemporal Multiresolution Associated to MRF Modelling for Motion Detection.- *Proc. ICIPA95*, Edinbourg, Ecosse, Juillet 1995, pp.158-162
[CI_Caplier95c] A. Caplier, F. Luthon. -A 3D space and time MRF-based Algorithm for Motion Detection.- *Proc. MCEA 95*, Grenoble, France, Septembre 1995, pp.147-152
[CI_Caplier95d] A. Caplier, F. Luthon. -A New Spatiotemporal Approach for Image Analysis. Application to Motion Detection.- *Proc. CAIP95*, Prague, République Tchèque, Septembre 1995, pp.246-253
[CI_Luthon94] F. Luthon, V. Popescu, A. Caplier -An MRF-based motion detection algorithm implemented on analog resistive network.- Proc. ECCV'94 , Stockholm, Suède, Mai 1994, pp.167-174
[CI_Luthon93] F. Luthon, A. Caplier. - Motion Detection and Segmentation in image sequences using Markov Random Fields Modelling.- *Proc. 4th Eurographics Animation and Simulation Workshop*, Barcelone, Espagne, Septembre 1993, pp.265-275

Conférences nationales

[CN_Hammal05] Z. Hammal., B. Bozkurt, L. Couvreur, D. Unay, A. Caplier, T. Dutoit - Classification bimodale d'expressions vocales- *GRETSI*, Septembre, Louvain La Neuve, Belgique, 2005
[CN_Benoit05] A. Benoit, A. Caplier – Interprétation des mouvements de la tête impliqué dans le processus de communication non verbale- *ORASIS05*, Clermont Ferrand, France, 2005
[CN_Hammal04a] Z. Hammal, A. Caplier, M. Rombaut - Classification des expressions faciales par la théorie de l'évidence- *12ème rencontres sur la logique floue et ses applications*, *LFA*, Nantes, France, Novembre 2004
[CN_Hammal04b] Z. Hammal, A. Caplier. - Analyse dynamique des transformations des traits du visage lors de la production d'une émotion – *Atelier Acquisition du geste humain par vision artificielle et applications, RFIA*, Toulouse, janvier 2004
[CN_Harasse04] S. Harasse, L. Bonnaud, A. Caplier, M. Desvignes. -Détection d'anomalies pour les caméras mobiles- *CORESA04*, France, may 2004
[CN_Hammal03] Z.Hammal, N. Eveno, A. Caplier, P.Y. Coulon - Extraction des traits caractéristiques du visage à l'aide de modèles paramétriques adaptés – *Colloque du Gretsi*, Paris, Sept. 2003
[CN_Bonnaud01] L. Bonnaud, A. Caplier, J.M. Chassery –Extraction rapide et robuste d'objets vidéo combinant une différence d'images et une image de référence réactualisée- *CORESA'01*, Dijon, France, juin 2001.
[CN_Caplier99] Caplier, N. Mottin -Détection automatique de lèvres dans un visage parlant. *CORESA'99*, Sophia-Antipolis, Juin1999, pp.239-246.

[CN_Caplier95] A.Caplier, F. Luthon -Algorithme markovien de détection de mouvement. Mise en oeuvre temps réel- *15ème Colloque du GRETSI*, Juans-les-Pins, France, Septembre 1995, pp.1033-1036.