



Université Blaise Pascal



CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

LASMEA - UMR 6602 UBP/CNRS

Laboratoire des Sciences et Matériaux pour l'Electronique et d'Automatique

Rapport sur le mémoire présenté par
M. Thomas BURGER
en vue de l'obtention du titre de Docteur
de l'Institut National Polytechnique de Grenoble

Les travaux de recherche entrepris par M. Thomas BURGER dans le cadre de sa thèse, se situent au confluent des deux domaines scientifiques complémentaires que sont la vision par ordinateur et la classification automatique. Ils concernent l'analyse des données délivrées par une caméra vidéo pour tendre vers l'interprétation automatique de la *Langue Française Parlée Complétée* (LPC). Cette étude s'inscrit dans le cadre du projet RNTS TELMA qui a pour objectif de donner accès à la téléphonie aux personnes malentendantes. Pour cela, il est proposé d'associer à un appareil téléphonique classique une caméra vidéo et un écran de visualisation ; la caméra permettant d'interpréter les gestes d'un locuteur mal-entendant codant le langage parlé complété dans un but de synthèse vocale, l'écran permettant d'animer un avatar afin de traduire les paroles d'un locuteur bien-entendant à destination d'un locuteur mal-entendant. C'est un projet particulièrement ambitieux qui nécessite, au sein d'un large consortium, le développement de très nombreuses briques algorithmiques complémentaires. Le candidat a focalisé son travail sur la phase délicate de la reconnaissance des gestes manuels participant à la *Langue Française Parlée Complétée*. Il s'agit, plus précisément, de reconnaître d'une part les configurations des doigts et d'autre part la position de la main vis-à-vis du visage.

Le premier chapitre du manuscrit permet à l'auteur d'introduire le contexte et quelques notions nécessaires à la justification des développements méthodologiques et théoriques à venir. Tout d'abord, les spécifications du langage parlé complété sont présentées. En effet, ce dernier est fondé d'une part sur la lecture labiale de chaque phonème prononcée et d'autre part sur une gestuelle permettant de différencier les phonèmes ambigus conduisant à des mouvements de lèvres identiques. La gestuelle sous-jacente est fondée sur huit configurations des doigts (codage des consonnes), couplée à cinq positions de la main relativement au visage (codage des voyelles). Face à la complexité de l'objectif visé et aux contraintes imposées au plan matériel (utilisation d'une seule caméra vidéo), des hypothèses simplificatrices sont proposées. D'une part, le codage gestuel devra être réalisé, autant que faire se peut, dans un plan parallèle au visage. D'autre part, afin de simplifier l'extraction automatique de la main, notamment en cas de superposition avec le visage, le codeur devra utiliser un gant coloré. Par la suite, des considérations temporelles sont introduites, avec notamment la notion d'images cibles. Il s'agit d'extraire les images où un geste est parfaitement réalisé en terme de configuration et de position. En pratique, il s'avère impossible aux codeurs de synchroniser totalement ces deux aspects. Le candidat préconise donc de séparer la notion d'images cibles de position et de configuration. Cette partie introductive se termine par la présentation de la structuration du reste du document.



24 Avenue des Landais 63177 AUBIERE Cedex (FRANCE)
Téléphone : 04-73-40-72-50 - Télécopie : 04-73-40-72-62 ☐ 04-73-40-73-40 ☐

Le chapitre suivant est centré sur les algorithmes "bas niveau" de traitement des images. Toutefois, préalablement M. Thomas BURGER présente les différents corpus de données exploités au sein de cette étude. Certains sont issus de l'observation, dans des conditions matériels jugées optimales, de codeurs certifiés. D'autres s'avèrent de moindre qualité de par une certaine imperfection du codage réalisé ou des conditions de prises de vue dégradées. Cette grande variété permettra de juger, au fil du document, de la robustesse des solutions préconisées. L'étape initiale concerne l'extraction automatique du gant dans les images. Cette dernière repose sur une phase d'apprentissage de la couleur de ce dernier dans l'espace YCbCr après désignation d'une zone d'intérêt dans la première image de la séquence. L'extraction de la zone correspondante dans les autres images, repose sur un test de similitude fondé sur la distance de Mahalanobis ainsi qu'une succession d'opérations permettant de combler les pixels manquants, de filtrer les fausses détections et d'extraire la plus grande composante connexe. La détermination du doigt pointeur est réalisée sur l'analyse du codage polaire du contour de la main en tenant compte de considérations morphologiques et des spécificités du langage parlé complété. Le positionnement des différentes zones de pointage (position de la main vis-à-vis du visage) repose sur l'utilisation d'un réseau de neurones (développé par un autre membre du consortium) qui couplé à un Perceptron Multi-couche fournit la position des yeux, du nez et de la bouche. Il est ainsi possible de positionner automatiquement, en lieu et en taille, cinq ellipses correspondantes aux zones de pointage du langage (coté du visage, pommette, bouche, menton, gorge). Afin de stabiliser temporairement les résultats ainsi obtenus, les différents paramètres des ellipses sont injectés dans un filtre de Kalman. La pertinence de chaque étape du schéma préconisé est validée à partir des différentes séquences du corpus.

Le troisième chapitre est consacré à la labellisation précoce. Il s'agit d'obtenir un étiquetage des images de la vidéo en images de transitions (geste intermédiaire entre deux codages) et images cibles (codage atteint). Comme indiqué préalablement, de par la désynchronisation des notions de *position* et de *configuration*, l'auteur préconise la recherche de deux étiquetages distincts. En ce qui concerne la position, l'étude se focalise simplement sur la trajectoire du centre de gravité de la main, cette dernière étant régularisée par un filtre de Kalman. Pour la configuration, afin de s'affranchir du mouvement global de la main, son image est tout d'abord normalisée (compensation de la translation du centre de gravité et recalage de l'axe principal d'inertie). Une configuration est jugée atteinte, si après normalisation, l'image de la main s'avère stable. Afin de traduire cette notion en mesure, l'auteur propose une approche bio-inspirée, intitulée *Filtre Rétinien Dédié*, qui permet de mettre en exergue les zones de mouvements inter-images. L'intégrale du signal ainsi obtenu fournit une mesure de la similitude de la configuration de la main sur une courte fenêtre temporelle. Munis de ces critères, les deux étiquetages souhaités sont obtenus par détection des plages de stabilité et désignation pour chacune d'elles d'une image cible. La confrontation au corpus de données permet de juger du bien fondée de l'approche proposée et également d'argumenter très finement les différentes causes d'échecs résiduels.

Le chapitre suivant qui concerne la reconnaissance du geste pour chaque image cible, apparaît comme le coeur de ce travail. Les deux aspects, position et configuration, sont abordés. En ce qui concerne la position, l'algorithme proposé est trivial dès lors que les pré-traitements décrits précédemment sont disponibles. La position retenue est celle correspondante à la zone désignée par le doigt pointeur. La reconnaissance de la configuration s'avère plus délicate. Tout d'abord, afin de renforcer la robustesse du processus, des traitements de nature géométrique sont proposés afin de supprimer le poignet et travailler uniquement sur le contour des doigts et de la paume de la main. Plusieurs attributs sont alors proposés afin de représenter la surface ainsi délimitée. Le premier d'entre eux est un indicateur global révélant la visibilité ou non du pouce dans l'image ; ce dernier permettant de scinder les configurations en deux familles disjointes. Par la suite, l'auteur propose l'utilisation de combinaison de moments d'inertie connu sous le vocable d'invariants de HU, ainsi que les descripteurs

71

de Fourier-Mellin ; ces derniers s'avérant, suite aux expérimentations menées, les plus performants. Le lecteur trouvera ensuite un état de l'art complet et très largement étayé concernant les méthodes de classification (*K plus proches voisins, réseaux de neurones, boosting, simplex, recuit-simulé, algorithmes génétiques, séparateurs à vastes marges ou SVM...*). La solution originale retenue est l'utilisation d'un banc de classifieurs SVM couplé à une technique de combinaison évidentielle (formalisme crédal ou des fonctions de croyance). Ceci permet une reconnaissance efficace de la configuration de la main dans les cas non ambigus mais également de gérer le doute dans le cas contraire. Les tests réalisés montrent la pertinence de l'approche proposée. Ce chapitre se termine sur des aspects plus théoriques montrant l'intérêt de la combinaison évidentielle en présence de classifieurs hétérogènes. Par la suite, le candidat explique comment étendre son utilisation au cas des classifieurs binaires non crédaux, unaires ou probabilistes.

Le cinquième chapitre est dédié à la fusion temporelle et multimodale des flux de configuration et de position du langage parlé complété. Devant la difficulté sous-jacente, seuls deux cas simplifiés sont traités qui concernent soit l'intégration temporelle, soit la combinaison multimodale. Le premier cas repose sur l'exploitation du corpus du LPC. La méthode proposée nécessite l'étiquetage des positions et des configurations sur toutes les images considérées comme transitoires. Ceci permet alors d'étendre les zones stables liées aux "images cibles" tant qu'il y a cohérence d'étiquetage. Enfin, l'identification des intersections temporelles non vides entre les deux ensembles permettent de retrouver les couples position-configuration malgré la désynchronisation des deux informations. La fusion multimodale est testée sur des séquences d'ASL (*American Sign Language*) où cohabitent des gestes de la main et de la tête. Cela permet la mise en oeuvre de la combinaison évidentielle de classifieurs binaires non crédaux évoquée au chapitre précédent. L'auteur indique qu'à sa connaissance il s'agit des premiers travaux traitant de la fusion des signes manuels et non-manuels dans le cadre de l'ASL. La solution préconisée permet un gain de l'ordre de 5% du taux de reconnaissance vis-à-vis de techniques classiques.

Le dernier chapitre est d'ordre purement théorique. Il propose une stratégie de décodage complet du langage parlé complété à partir des trois flux position, configuration et labial (lecture des lèvres). Après quelques considérations d'adaptation des techniques évoquées précédemment dans le cadre de la reconnaissance labiale, M. Thomas BURGER propose une méthode de synchronisation des trois modalités reposant sur une modélisation par un processus de Markov à sauts caché et une résolution fondée sur l'algorithme de Viterbi. Un exemple de la machine à état sous-jacente est donné. Ce chapitre théorique laisse entrevoir de nombreuses suites à ce travail.

D'une manière générale le manuscrit est clair et très agréable à lire. Les développements méthodologiques sont bien introduits et rigoureusement menés ; ce qui permettra au lecteur intéressé d'aborder ces étapes dans d'excellentes conditions. La valeur intrinsèque du travail mené est indéniable comme le prouvent les nombreuses publications de l'auteur dans des revues ou des conférences de renom. En conclusion, je considère que le mémoire présenté par M. Thomas BURGER est d'un excellent niveau scientifique ; les techniques proposées de reconnaissance des gestes manuels dans le cadre de la *Langue Française Parlée Complétée* laissant entrevoir des perspectives applicatives particulièrement intéressantes. Pour toutes ces raisons, j'émet un avis sans réserve quant à la soutenance de cette thèse en vue de l'obtention du titre de docteur de l'Institut National Polytechnique de Grenoble.

Fait à Aubière, le 14 octobre 2007


Michel DHOME
Directeur de Recherche au CNRS