

---

**Rapport sur le mémoire de thèse intitulé :**

*Reconnaissance automatique des gestes de la Langue Française Parlée Complétée*

**présenté par :** Monsieur Thomas BURGER

**pour :** l'obtention du grade de docteur de l'INPG.

**Rapport établi par :** Olivier COLOT

Professeur à l'Université des Sciences et Technologies de Lille  
Directeur adjoint du LAGIS – UMR CNRS 8146.

---

**1. Point de vue général**

Monsieur Thomas BURGER a effectué ses travaux de recherche au sein du laboratoire GIPSA-Lab – UMR CNRS 5083 de l'INPG et de France Télécom R&D dans le cadre d'une bourse CIFRE sous la direction scientifique d'Alice CAPLIER. Le travail de Thomas BURGER s'inscrit dans le cadre du projet RNTS – TELMA.

Le mémoire de 306 pages comporte sept chapitres, une bibliographie de 176 références (plus les références de l'auteur), trois annexes permettant d'alléger le texte principal de certains développements théoriques ou pratiques, une liste des tableaux, une liste des figures, un glossaire et une table des matières. La qualité rédactionnelle du mémoire est très bonne rendant sa lecture très agréable.

Dans sa thèse, Monsieur Thomas BURGER propose une architecture de traitements de séquences vidéo en couleur destinée à la reconnaissance automatique des gestes de la Langue française Parlée Complétée (LPC) ou code LPC. Le travail du candidat s'inscrit dans un projet de grande ampleur, impliquant différents acteurs sur les différentes facettes que comporte ce projet. Thomas BURGER s'est focalisé sur certains aspects de TELMA et propose des solutions pertinentes et originales pour répondre de manière adaptée à certains problèmes clés.

Dans le mémoire à la disposition du rapporteur, c'est donc un système de traitement complet qui est proposé, s'étendant de l'acquisition de séquences d'images à la reconnaissance de gestes. Il couvre ainsi un large ensemble de domaines incluant le traitement d'images, le filtrage, la fusion de données et la reconnaissance de formes.

On notera cependant que de nombreuses références bibliographiques mériteraient d'être complétées (des numéros de pages, de volumes sont absents).

**2. Analyse détaillée**

Le premier chapitre d'introduction du mémoire comporte 16 pages. Thomas BURGER expose le contexte dans lequel s'inscrit son travail de thèse et expose l'objectif visé et la finalité de ses recherches. Ainsi, il positionne le champ d'étude de son travail dans le cadre du projet RNTS – TELMA, et précise le positionnement de ses recherches tant dans son contexte applicatif que dans son contexte scientifique. Au sein du vaste projet TELMA, Thomas BURGER s'est focalisé sur la problématique de l'acquisition des gestes signifiants du point de vue de la LPC (configurations et positions) ainsi que sur la problématique de leur interprétation. A terme, il s'agit, à partir des codes reconnus par le processus de traitement, de piloter un

système de synthèse vocale devant transmettre un message par liaison téléphonique, message qui a la réception animera un avatar reproduisant les codes (gestes) qui seront présentés à un utilisateur final malentendant. Il expose les grandes étapes qu'il développe dans son mémoire en précisant les liens qui unissent chacune de celles-ci.

Le deuxième chapitre (22 pages) intitulé « Description des gestes du LPC » présente les spécificités et les difficultés posées par la reconnaissance automatique de la LPC. L'interprétation des gestes de la LPC doit reposer sur l'obtention d'images cibles (IC) dans un flux vidéo présentant des configurations et des positions de la main les plus nets possibles pour éviter toute ambiguïté ou impossibilité de reconnaissance. Cependant, une certaine souplesse doit être intégrée dans le système de reconnaissance automatique afin de ne pas imposer des contraintes trop fortes qui rendraient l'exploitation du système difficile ou quasi impossible, en particulier au regard de l'usage final visé.

La fin de ce chapitre rappelle l'organisation des chapitres suivants et propose un graphique très judicieux pour présenter de manière synthétique le thème principal de chacun des chapitres suivants et leurs liens, ce qui permet d'avoir un aperçu très synthétique et structuré de l'organisation du mémoire.

Le troisième chapitre (43 pages) intitulé « analyse et segmentation des images » est consacré à la segmentation de zones d'intérêt dans l'image (visage, main) afin de permettre le décodage ultérieur des images de position et de configuration. La première étape consiste à segmenter la main au moyen d'un triple seuillage. Les trois seuils proposés pour mener à bien cette segmentation sont déterminés de manière empirique. On peut s'interroger sur la robustesse du processus de segmentation vis-à-vis des valeurs de seuils choisies. Une discussion plus approfondie aurait été intéressante.

Ensuite, l'identification du doigt pointeur est effectuée selon un processus décomposé en huit étapes.

Dans un troisième temps, le candidat exploite la technique CFF (*Convolutional Face Finder*) et C3F (*Convolutional Face and Feature Finder*) de Garcia *et al.* afin de définir la zone du visage pointée dont la détermination est essentielle pour déterminer la configuration de la main et la position (zone pointée du visage ou à côté du visage) afin de reconnaître le code LPC et donc le sens lié au couple configuration – position du geste du codeur.

Les résultats produits montrent les très bonnes performances des traitements proposés sur les images de vidéo analysées.

Le quatrième chapitre (23 pages) intitulé « labellisation précoce » est consacré à la séparation des images cibles (destinées à la reconnaissance de la position et de la configuration) des images dites de transition. Pour ce faire, le candidat propose de mener une segmentation des séquences d'images en s'appuyant sur des informations cinématiques propres à permettre la distinction entre des zones de stabilité et des zones de transition. Les gestes à reconnaître comportent à la fois une information de position (un doigt dénommé « pointeur » pointant sur une zone spécifique parmi cinq possibles) et une information de configuration (forme de la main et en particulier les doigts participant aux gestes). L'une des principales difficultés réside dans la désynchronisation pouvant exister entre l'instant correspondant à un pointage stable d'une zone signifiante et l'instant de présence ou stabilisation d'une configuration exploitable. Thomas BURGER propose d'analyser séparément la position de la configuration.

Tout d'abord, une analyse globale du mouvement de la main lors du changement de position est effectuée. Celle-ci se fonde sur la trajectoire du centre de gravité de la main (dont l'auteur explique que l'expérience montre qu'elle est plus robuste que le centre de la paume). Ensuite, une analyse de la déformation de la main lors de changements de configuration est menée. Celle-ci s'appuie sur une évaluation de la quantité de mouvement et dont l'analyse doit permettre de détecter les zones de stabilité de celles présentant des transitions. Le processus mis en œuvre est appelé Filtre Rétinien Dédié (FRD) et se décompose en six étapes :

1. une segmentation de la main telle que proposée au chapitre 3 ;
2. une détection de contours permettant de mettre en valeur les doigts ;
3. une opération de pondération appliquée sur l'image des contours ;

4. un lissage par des filtres de type gaussien ;
5. un filtre IPL (proposé par A. Benoît dans sa thèse) dont la sortie renseigne sur la quantité de mouvement ;
6. et enfin, une opération de sommation permettant d'obtenir un indicateur global de quantité de mouvement.

La réponse du FRD est lissée par deux convolutions successives à l'aide de masques approximant des filtres gaussiens et constitue alors un signal de quantité de mouvement sur lequel des seuillages sont effectués afin de localiser des zones de stabilité et des zones de transition. Le choix des seuils est discuté. L'auteur préconise d'utiliser des seuils sélectifs sans nuire pour autant à la robustesse du processus de d'identification des zones d'information recherchées (zones de stabilité, zones de transition). Les résultats produits sur les vidéos testées montrent la pertinence de la méthodologie proposée. Néanmoins, Thomas BURGER souligne que des améliorations ou des modifications dans les traitements sont possibles.

Une étude et une discussion plus approfondies auraient été bienvenues sur certaines étapes du traitement, comme par exemple le choix de la matrice de pondération appliquée sur l'image des contours ou les opérations de lissage par masques gaussiens.

Le cinquième chapitre (59 pages) intitulé « reconnaissance du geste statique » est, avec le chapitre suivant, au cœur de la contribution du candidat. Deux phases essentielles sont distinguées : la reconnaissance de la position (identification de la zone pointée par le codeur parmi un ensemble de quatre zones prédéfinies) puis ensuite la reconnaissance de la configuration des doigts de la main afin de permettre la reconnaissance complète du geste.

La première étape du processus de reconnaissance (reconnaissance de la position) est destinée à associer le doigt identifié comme pointeur à l'une des cinq zones de pointage prédéfinies sur les images cibles à traiter. Les différentes zones sont circonscrites par des cercles ou des ellipses et sont localisées par rapport au visage du codeur à partir de considérations morphologiques telles que proposées dans la littérature. Cette phase est relativement simple dans le principe. Les résultats produits par l'auteur montrent le bon comportement de la méthode proposée. Ceux-ci sont analysés et commentés de manière pertinente.

La phase de reconnaissance de la configuration est, quant à elle, beaucoup plus complexe. Elle nécessite la mise en œuvre de processus de discrimination élaborés afin d'atteindre des performances compatibles avec les objectifs recherchés. Pour y parvenir, une étude de différents descripteurs de forme est menée et Thomas BURGER retient un ensemble de trente-trois descripteurs de Fourier-Mellin et un descripteur associé à la présence ou non du pouce dans la configuration à analyser.

Ensuite, l'auteur propose de répondre à la problématique de reconnaissance des configurations en utilisant un processus de combinaison crédibiliste d'un banc de SVM (C-SVM) dont la fonction noyau est une sigmoïde..

Les fonctions de croyance utilisées dans le processus de combinaison sont construites à partir de sous-ensembles flous définis à partir d'une information de distance à un hyperplan séparateur. Le point central réside dans la modélisation d'une zone d'hésitation (incertitude) dont la définition par une fonction d'appartenance aurait méritée d'être plus discutée – en particulier – quant à son influence sur le processus de décision ultérieur.

Les SVM travaillant sur des cadres de discernement différents (chaque SVM se prononce sur un groupe restreint de configurations), il est nécessaire de transformer les fonctions de croyance définies pour chaque SVM sur un cadre commun. L'argumentation sur la transformation appliquée aux fonctions de croyance élémentaires construites pour chaque SVM pour les exprimer sur un cadre de discernement commun n'est pas très claire. Je ne suis d'ailleurs pas convaincu qu'il s'agisse réellement d'un raffinement (au sens commun en théorie des fonctions de croyance) comme le suggère l'auteur.

Les résultats obtenus sur différents corpus montrent l'intérêt du processus de reconnaissance proposé par. L'influence des paramètres de réglage des SVM sur les résultats de reconnaissance est étudiée. Les taux de reconnaissance en situation mono-codeur et multi-codeur montrent des performances tout à fait satisfaisantes.

Enfin dans une dernière partie de ce chapitre, Thomas BURGER propose de généraliser le principe de la combinaison évidentielle à d'autres types de classificateurs qu'ils soient binaires non crédaux, unaires ou probabilistes. Pour ce dernier cas, il propose une transformation crédale permettant d'exprimer une probabilité en une fonction de croyance. De fait, l'auteur définit une transformée inverse de la transformée pignistique.

Dans le sixième chapitre (33 pages) intitulée « interprétation phonémique », un double objectif est recherché : une fusion temporelle et une fusion multimodale des flux de configurations et de positions de la LPC. Si l'aspect temporel est effectivement traité pour la LPC, l'aspect multimodal est, quant à lui, étudié sur une autre application : la langue des signes américaine (ASL).

Pour l'aspect temporel, Thomas BURGER propose une méthode permettant d'associer une image cible de configuration avec une image cible de position en résolvant les problèmes liés à la désynchronisation et au décalage entre la réalisation des configurations et l'atteinte des positions significatives dans le flux vidéo. La méthode de correction des désynchronisations et décalages proposée se décompose en quatre étapes. L'idée est d'effectuer le rallongement des plages de stabilité de configurations et de positions (obtenues par la méthode exposée au chapitre 3) afin d'assurer l'existence d'intersections non vides entre les plages de stabilité des deux informations (configuration et position).

La seconde partie du chapitre est dédiée à la combinaison multimodale, dont l'application a été faite sur des vidéos de l'ASL (dans le cadre d'un échange scientifique avec un laboratoire turc) et non sur la LPC. Les informations à traiter dans le cas de l'ASL sont multimodales (informations manuelles et non manuelles).

Le point fort de cette partie réside principalement dans la définition d'une transformée pignistique partielle (PPT) tout à fait innovante dans sa conception (des éléments de justification théorique sont proposés dans l'annexe B du document). L'objectif de cette approche est de converger vers une décision pertinente en situation d'hésitation entre différentes solutions. Le principe consiste à examiner différents cadres de décision dont les structures sont liées à un paramètre  $\gamma$ . Typiquement, la cardinalité des ensembles solutions est lié à ce paramètre. La PPT se comporte par ailleurs comme la transformée pignistique classique de Smets lorsqu'elle considère uniquement des solutions singletons.

Pour l'application en reconnaissance de signes de l'ASL, le principe consiste à construire dans un premier temps des fonctions de croyance à partir des vraisemblances issues de modèles de Markov cachés (HMM) modélisant des informations manuelles et non manuelles. Ces croyances sont ensuite combinées puis traitées par la PPT. Ceci constitue la première étape de la classification. Si la PPT ne met pas en évidence d'hésitation, le processus est terminé. Dans le cas contraire, une seconde classification est menée pour affiner l'analyse en étudiant plus spécifiquement les sous-ensembles (clusters) caractérisant l'hésitation. Cette dernière étape utilise uniquement les informations non manuelles et permet de finalement décider par maximum de vraisemblance parmi l'ensemble des solutions candidates. Les résultats fournis en termes de taux de reconnaissance montrent que la méthode proposée produit des résultats supérieurs par rapport à des approches plus classiques.

Le septième chapitre (22 pages) intitulé « vers la fusion main/lèvre » pose les bases pour les futurs développements du travail de recherche avec pour objectif la finalisation d'un système de reconnaissance complet du code LPC.

Dans ce chapitre, Thomas BURGER expose les difficultés liées à la segmentation des lèvres (en particulier dans le contexte d'acquisition des images imposé par l'application finale) et l'extraction de caractéristiques propres à renseigner quant au signe labial émis par un codeur (problème de reconnaissance).

Le candidat expose également toute la problématique liée à la fusion multimodale permettant d'associer les gestes manuels (configurations et positions) aux informations labiales, en mettant en exergue les difficultés liées à la désynchronisation des messages gestuels et labiaux, ce qui rend cette fusion délicate.

Ce chapitre constitue essentiellement une « feuille de route » exposant une stratégie possible pour aboutir à une reconnaissance complète des codes de la LPC tenant compte de son aspect multimodal et des difficultés liées au décalage temporel des différents flux d'informations impliqués. Les pistes évoquées dans

ce chapitre montrent que l'auteur a déjà des idées précises pour résoudre les problèmes de reconnaissance intégrant les informations labiales et les gestes du code LPC. Il y a donc là l'esquisse d'un projet de recherche pour l'avenir.

Le dernier chapitre de conclusion (4 pages) reprend les éléments essentiels du travail et ouvre sur quelques perspectives à donner pour apporter des améliorations sur certaines parties des traitements proposés et des éléments reprenant les pistes évoquées au chapitre 7 pour prolonger les travaux et compléter le processus de reconnaissance automatique de la LPC. L'auteur propose également de prolonger certains aspects théoriques exposés dans son mémoire.

Enfin, trois annexes sont proposées : la première concerne les fonctions de croyance, la deuxième propose des justifications théoriques de la PPT et la dernière comporte des compléments sur certaines méthodes issues de la littérature que le candidat a utilisées à certaines étapes de son processus (CNN, CFF,...).

### 3. Avis du rapporteur

Le travail de Thomas BURGER est conséquent et de qualité tant du point de vue théorique que pratique. Il propose ainsi des solutions performantes en termes de reconnaissance de gestes de la LPC ou de l'ASL et les résultats produits dans le document sont convaincants.

Thomas BURGER démontre une bonne maîtrise de l'ensemble des domaines concernés par son travail. Certaines contributions théoriques présentées dans le document sont originales et dignes d'intérêt. Notons également que le candidat démontre une réelle capacité à prendre de la hauteur par rapport à son travail, ce qui démontre une maturité certaine. Certains de ses travaux ont été menés dans le cadre d'une collaboration avec des partenaires étrangers, ce qui démontre également son ouverture d'esprit et des qualités pour le travail en équipe.

Il est à noter que les travaux du candidat ont donné lieu à un article accepté pour publication dans une revue internationale, à six communications dans des conférences internationales et à trois soumissions d'articles dans des revues internationales.

Pour toutes ces raisons, j'émet un avis très favorable à la soutenance publique de la thèse de Monsieur Thomas BURGER devant la commission d'examen en vue de l'obtention du grade de docteur de l'INPG.

A Villeneuve d'Ascq, le 10 octobre 2007.



Professeur Olivier COLOT  
Université des Sciences & Technologies de Lille  
Directeur adjoint du LAGIS – UMR CNRS 8146