Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia.

Pierre Gacon

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité - Signal, Image, Parole, Télécoms

préparée au Laboratoire des Images et Signaux (LIS) dans le cadre de l'Ecole Doctorale
Electrotechnique, Electronique, Automatique, Traitement du Signal

présentée et soutenue publiquement par

Pierre Gacon

le 19 juillet 2006

Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia.

Directeurs de thèse : Pierre-Yves COULON et Gérard BAILLY

JURY

M. James CROWLEY	Professeur INPG	Président
M. Marinette REVENU	Professeur ENSICAEN	Rapporteur
M. Frank DAVOINE	Chargé de recherche CNRS	Rapporteur
M. Pierre-Yves COULON	Professeur INPG	Directeur de thèse
M. Gérard BAILLY	Directeur de Recherche CNRS	Co-directeur

REMERCIEMENTS

Tout d'abord, je remercie Pierre-Yves pour son encadrement et sa présence tout au long de notre collaboration, depuis mon stage de DEA jusqu'à la conclusion de cette thèse.

Je remercie ensuite Gérard pour sa rigueur scientifique et les nombreuses idées et pistes qu'il m'aura apporté.

Enfin, je remercie les membres du jury :

- James Crowley pour avoir accepté de présider à ma soutenance et pour les très intéressants points que ses questions ont soulevés.

- Marinette Revenu et Frank Davoine pour avoir accepté d'expertiser le présent rapport et pour leurs remarques qui auront contribué à la qualité de la version finale.

En outre, je veux remercier et associer à cette thèse Zakia, Corentin et Mickaël qui auront été mes compagnons de labeur tout au long de ses quatre années et demi passées au LIS.

TABLE DES MATIERES

TABLE DES FIGURES

Chapitre 2

Annexe 1

Annexe 2

LISTE DES TABLEAUX

Annexe 1

NOTATIONS

Convention :

M, **m** : Matrices en gras.
V, ***v*** : Vecteurs en gras italique.
S, s : Scalaires en italique.

Principales Notations :

Modèle de forme (chapitre 2 et 3) :
$\boldsymbol{X}$, $\overline{\boldsymbol{X}}$, $\boldsymbol{x}$, $\boldsymbol{p_{x,j}}$, $\lambda_{x,j}$, $\mathbf{P}_x$, nb_x : Vecteur de donnée, vecteur moyen, vecteur de paramètres du modèle, vecteurs propres, valeurs propres, matrice de vecteurs propres, nombre de vecteurs propres.

Modèle d'apparence (chapitre 2) :
$\boldsymbol{A}$, $\overline{\boldsymbol{A}}$, $\boldsymbol{a}$, $\boldsymbol{p_{a,j}}$, $\lambda_{a,j}$, $\mathbf{P}_a$, nb_a : Vecteur de donnée, vecteur moyen, vecteur de paramètres du modèle, vecteurs propres, valeurs propres, matrice de vecteurs propres, nombre de vecteurs propres.

Modèle d'apparence statique (chapitre 3) :
$\boldsymbol{S}$, $\overline{\boldsymbol{S}}$, $\boldsymbol{s}$, $\boldsymbol{p_{s,j}}$, $\lambda_{s,j}$, $\mathbf{P}_s$, nb_s : Vecteur de donnée, vecteur moyen, vecteur de paramètres du modèle, vecteurs propres, valeurs propres, matrice de vecteurs propres, nombre de vecteurs propres.

Modèle d'apparence dynamique (chapitre 3) :
$\boldsymbol{D}$, $\overline{\boldsymbol{D}}$, $\boldsymbol{d}$, $\boldsymbol{p_{d,j}}$, $\lambda_{d,j}$, $\mathbf{P}_d$, nb_d : Vecteur de donnée, vecteur moyen, vecteur de paramètres du modèle, vecteurs propres, valeurs propres, matrice de vecteurs propres, nombre de vecteurs propres.

Modèle combiné (chapitre 2 forme/apparence, chapitre 3 forme/apparence dynamique) :
$\boldsymbol{C}$, $\overline{\boldsymbol{C}}$, $\boldsymbol{c}$, $\boldsymbol{p_{c,j}}$, $\lambda_{c,j}$, $\mathbf{P}_c$, nb_c, W_e : Vecteur de donnée, vecteur moyen, vecteur de paramètres du modèle, vecteurs propres, valeurs propres, matrice de vecteurs propres, nombre de vecteurs propres, coefficient de normalisation.

INTRODUCTION

Cadre de la thèse

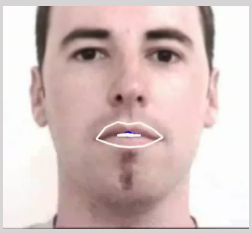
Dans le domaine très large du traitement de l'image, la vision par ordinateur a pris une part de plus en plus importante au fur et à mesure que les opérations de base et techniques de segmentation d'image se perfectionnaient.

L'analyse d'image ne se limite plus à segmenter une image en régions ou à en détecter les contours, l'ordinateur doit à présent être capable d'exploiter ces informations pour donner un sens aux images, de la même façon que l'homme est capable d'interpréter son environnement grâce à sa vue. Les champs applicatifs sont très diversifiés, on peut citer à titre d'exemples la télésurveillance, la télédétection ou la biométrie.

Parmi toutes les images pouvant servir de sujet d'étude, celles représentant le visage humain sont devenues un centre d'intérêt particulièrement fort du domaine de la vision par ordinateur. Le visage fournit en effet quantité d'informations que le cerveau humain peut relever et interpréter de façon naturelle : la plus évidente est l'identité de la personne mais on peut relever également l'état émotionnel, la direction du regard et donc le centre d'attention, etc…

Si les traits du visage sont donc remplis de sens, la zone qui en contient le plus est celle de la bouche car elle contient des informations caractéristiques liées à la parole et donc à la communication entre être humains (voir figure 0.1).

La volonté d'obtenir des interfaces établissant des rapports de plus en plus naturels entre l'humain et la machine fournit donc un terrain d'application pratique privilégié pour les nombreux algorithmes de segmentation du visage et des lèvres développées au cours des vingt dernières années.



→
- Animation de tête parlante virtuelle
- Reconnaissance de la parole
- Analyse d'émotion
- Identification du locuteur

Figure 0.1 :
Exemples d'application de l'analyse labiale.
L'information obtenue à partir d'une segmentation des lèvres effectuée par un algorithme dédié peut être utilisée pour différentes applications.

Sur le pôle scientifique grenoblois, les travaux d'analyse labiale remontent aux années 80 et aux travaux de [Lallouache, 1991] à l'Institut de la Communication Parlée (ICP) dans lesquels un maquillage bleu était utilisé pour capturer le mouvement des lèvres. Depuis, cette activité s'est largement développée et plusieurs travaux ont été menés au Laboratoire des Images et des Signaux (LIS) comme par exemple ceux de Liévin ([Liévin, 2000]) qui utilisait des techniques utilisant des informations colorimétriques et de mouvements pour segmenter des indices faciaux. Dans [Delmas, 2000], les contours actifs ont été mis en oeuvre pour détecter les contours des lèvres. Enfin, dans [Eveno, 2003], l'auteur a, quant à lui, eu recours à des modèles analytiques déformables pour accomplir la même tâche.

Contrairement à ces précédentes approches, cette thèse met en oeuvre les techniques de modélisation statistique popularisée au cours de la dernière décennie dans la communauté de l'analyse labiale : les Modèles de Formes Actifs (ASM) et les Modèles d'Apparence Actifs (AAM), techniques dans lesquelles l'ICP a développé une solide expertise, notamment en les appliquant à l'analyse-synthèse de visages parlants pilotés par des modèles articulatoires ([Odisio, 2005]).

Ces techniques nécessitent d'avoir des bases d'images où les points caractéristiques des lèvres ont été manuellement annotés. Le principe de base des modèles actifs consiste ensuite à procéder à des analyses en composantes principales sur ces données d'apprentissage. Cela permet d'obtenir un modèle contrôlé par un nombre limité de paramètres et capable de générer n'importe-quelle configuration labiale proche de celles de la base de données. Cette approche évite donc les étapes de paramétrisation de la modélisation qui sont en général empiriques.

Dans ce travail, nous nous sommes particulièrement intéressés à la zone de l'intérieur de la bouche (qui n'avait pas été abordée par la travaux antérieures menés au LIS) et aux problèmes posés par ses non-linéarités (ouverture/fermeture de la bouche, présence/absence des dents). En outre, la méthode devait être capable de traiter avec robustesse des images de locuteurs variables acquises dans des conditions diverses.

Le LIS et l'ICP ont participé ensemble au projet RNRT TEMPOVALSE qui visait à mettre en œuvre les techniques de segmentation labiale dans une application concrète de terminal audiovisuel portatif ([Bailly, 2003]). Le présent travail a ainsi été réalisé dans la continuité de la collaboration entre les deux laboratoires.

Le chapitre 1 présente les applications de l'analyse labiale, et donc les débouchés possibles de la présente étude, ainsi qu'un état de l'art des diverses approches envisagées. Un accent particulier est mis sur la méthodologie des ASMs et AAMs et les cadres dans lesquels ils ont été appliqués.

Les chapitres 2 et 3 de ce rapport reprennent la structure de développement méthodologique qui a été effectivement suivi au cours de ce travail. Un algorithme de segmentation peut être qualifié selon sa précision, sa robustesse, son coût en calcul ou son applicabilité. Cette dernière caractéristique correspond au cadre dans lequel l'algorithme

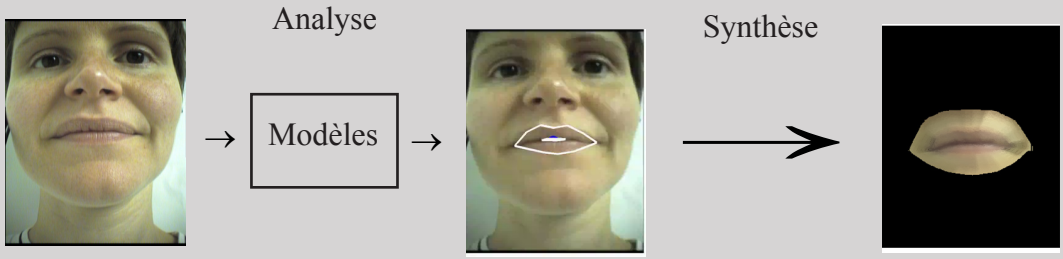


Figure 0.2 :
Schéma d'analyse/synthèse de la zone labiale présenté dans ce rapport.
Après acquisition d'une séquence vidéo inconnue, la partie analyse correspond à la segmentation du contour des lèvres grâce au recours à différents modèles statistiques. La partie synthèse correspond à la génération d'une séquence synthétique fidèle à l'originale et restituant ainsi l'information.

peut être exécuté tout en conservant sa précision et sa robustesse. Certaines méthodes ne s'appliquent par exemple qu'à un unique locuteur pris dans un seul type de condition, tandis que d'autres se veulent universels et pouvant s'appliquer à tout individu sans contraintes.

Les algorithmes ont donc d'abord été développés et testés dans un cadre mono-locuteur (chapitre 2) avant que leur champ d'application ne soit étendu à un cas multi-locuteurs (chapitre 3) avec les modifications nécessaires pour conserver la robustesse et la précision. Le chapitre 2 introduit en outre une fonction de coût faisant intervenir une association non-linéaire entre la forme et une description locale de l'apparence. Le chapitre 3 présente quant à lui une séparation de l'apparence entre des composantes statique et dynamique dans l'objectif de ramener le problème multi-locuteurs au cadre mono-locuteur du chapitre précédent et donc de rendre les techniques développées au chapitre 2 transposables à un cadre plus large. En outre, deux modèles statistiques servant de prétraitement sont également présentés : le premier détecte les commissures des lèvres tandis que le second classifie les pixels comme étant de lèvre ou de peau.

Enfin, le chapitre 4 propose l'évaluation des performances de la méthode proposée, selon des critères quantitatifs et qualitatifs. L'évaluation qualitative sera effectuée dans le cadre d'un schéma d'analyse/synthèse où la segmentation des lèvres permet de générer un avatar virtuel de la bouche du locuteur pouvant être utilisé pour améliorer la compréhension de la parole en milieu bruité (voir figure 0.2).

CHAPITRE 1

Applications et Méthodologies de l'Analyse Labiale

Après une présentation de la nature bimodale de la parole qui justifie en grande partie l'analyse labiale, nous verrons quelques domaines d'applications possibles du travail présenté dans ce rapport de thèse : l'animation de clone, la reconnaissance automatique de parole et la reconnaissance d'émotions. Cette présentation du champ applicatif n'est cependant pas exhaustive, l'on pourrait rajouter, entre autre, la biométrie par analyse statique ou dynamique des traits du visage (les lèvres n'étant alors qu'un des points d'intérêt).

Après la revue des applications justifiant les techniques d'analyse labiale, nous verrons un état de l'art des différentes méthodologies utilisées pour accomplir cette tâche : les approches pixels, les approches contours utilisant des modèles des lèvres (analytiques ou statistiques) et la famille des modèles statistiques de forme et d'apparence à laquelle appartient le présent travail.

1.1 Applications de l'analyse labiale

1.1.1 La nature bimodale de la parole

L'une des illustrations les plus familières de l'aspect bimodal de la parole est la faculté des personnes malentendantes à comprendre le discours d'un interlocuteur en s'aidant des mouvements des lèvres. Cette faculté est en fait universelle et, entendants ou sourds, nous utilisons la cohérence audiovisuelle de manière inconsciente.

De nombreuses études ont à présent clairement établi et formalisé cet aspect multimodal, le cerveau intégrant, lorsqu'elles sont disponibles, les informations auditives et visuelles pour interpréter un son.

Des expérimentations ont démontré l'importance de cette fusion d'informations comme dans [McGurk, 1976], où des sujets ont été mis en présence de sources auditives et visuelles contradictoires conduisant à la perception d'un troisième son. Cette constatation a été baptisée « effet McGurk » : par exemple, si une personne est confrontée au phonème 'ba' et au visème 'ga', elle aura l'impression d'entendre 'da', l'information visuelle conduisant donc dans un tel cas à une mauvaise compréhension de la parole.

La parole « audiovisuelle » peut être partitionnée du point de vue perceptif en unités élémentaires baptisées phonèmes (au nombre de 48 dans le cas de l'anglais, [Rabiner,

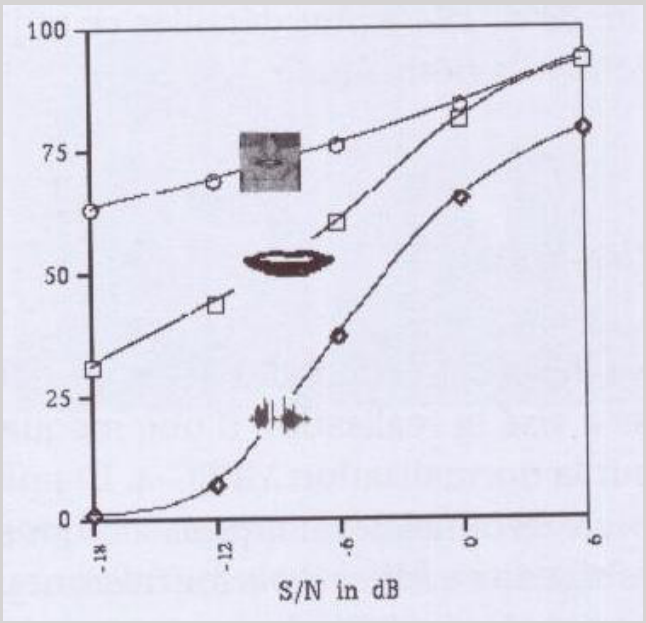


Figure 1.1 :
Taux de compréhension en milieu bruité de logatomes pour trois modalités: visage+audio, lèvre+audio, audio seul, d'après [le Goff, 1995].

1993]). En pratique un même phonème peut correspondre à des réalisations sonores différentes en fonction des locuteurs et des conditions extérieures : les allophones. On a également défini une unité élémentaires d'information visuelle (ou allophone visuel) : le visème (l'anglais ayant 9 familles de visèmes, [Dodd, 1987]).

Les allophones sonores et visuels apportent des informations complémentaires et avoir accès aux deux lève les ambiguités. L'information visuelle aide donc la compréhension d'un discours par un auditeur. Par exemple les sons 'b' et 'd', proches d'un point de vue auditif, sont facilement différentiables grâce à la bimodalité.

Cette cohérence audiovisuelle permet notamment de mieux comprendre la parole en présence de bruit sur le canal auditif. Deux personnes discutant au milieu d'une foule auront ainsi spontanément tendance à regarder les lèvres de leur interlocuteur afin d'obtenir un rehaussement « naturel » du signal en augmentant l'information ce qui conduit à un gain audiovisuel de 11dB en rapport signal/bruit ([MacLeod, 1987], [Summerfield, 1989]). De même, l'information visuelle est utile lorsque le contexte linguistique (la langue, l'accent ou le niveau de langage) n'est pas familier ([Reisberg, 1987], [Reisberg, 1987+]).

Cet effet a été démontré par de nombreuses études, par exemple dans [Le Goff, 1995], nous avons des scores de compréhension de logatomes pour trois modalités différentes : le son seul, le son et les lèvres, et le son et le visage entier (figure 1.1). On remarquera que même dans le cas d'un signal clair non bruité on a un rehaussement d'environ 20% de la compréhension apporté par la modalité visuelle.

Si on constate une augmentation de la compréhension des deux modalités incluant des informations visuelles même avec un bruit faible, plus le bruit sera fort par rapport au signal audio et plus le gain en compréhension sera important. En outre, la différence entre les modalités « lèvres seules » et « visage entier » suggère que le cerveau intègre d'autres indices visuels que le seul mouvement des lèvres.

1.1.2 Têtes parlantes virtuelles : communication audio-visuelle et autres applications multimédia

L'apport de l'information visuelle pour deux locuteurs ayant été démontré par diverses études, la possibilité d'utiliser cette modalité dans des systèmes de visiophonie a été naturellement envisagée. Jusqu'à une époque récente les débits limités des lignes téléphoniques ou Internet rendaient impossible ou presque l'envoi en temps réel du visage des locuteurs. Or, si la cadence d'images par seconde descend en dessous de la douzaine, le bénéfice apporté disparaît que cela soit en vidéo naturelle ([Vitkovitch, 1994], [Gagné, 1997]) ou synthétique ([Pandzic, 1999]).

Divers projets ont donc proposé de n'envoyer que des paramètres codant la forme des lèvres, le visage du locuteur étant remplacé par un avatar ou clone virtuel qui pourrait alors être animé en temps réel suivant les mouvements réels ce qui ne nécessiterait plus que l'envoi de quelques bits d'informations.

Figure 1.2 :
Télécommunication en utilisant le dispositif TEMPOVALSE.
De gauche à droite et de haut en bas: locuteurs équipés, casque capturant l'image, terminal de réception, image réceptionnée, schéma de principe (d'après [Bailly, 2003]).

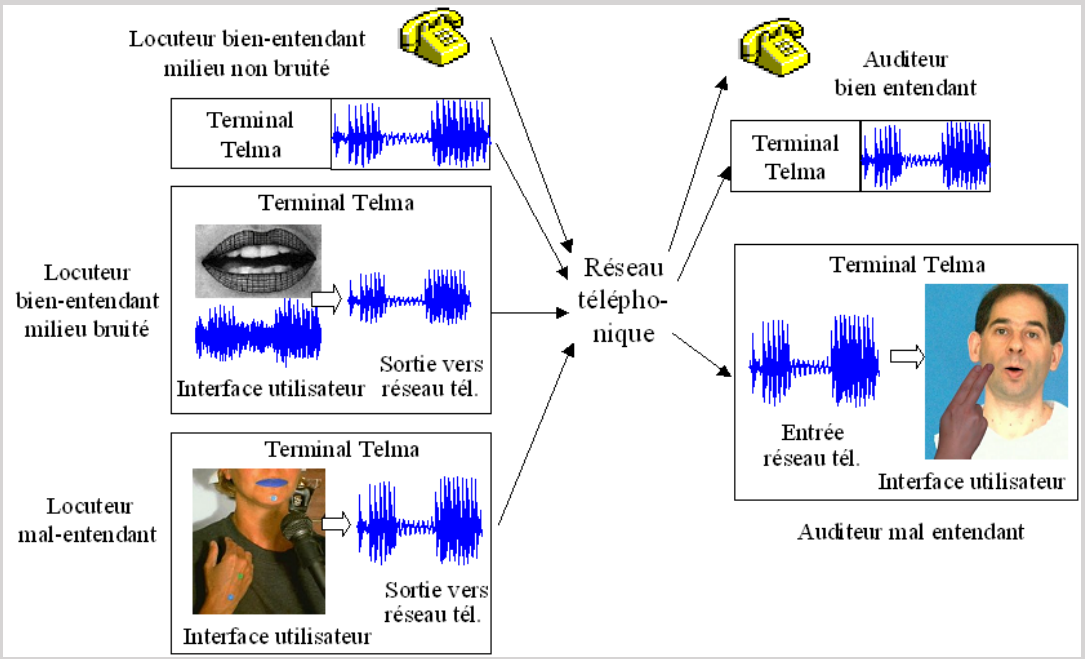


Figure 1.3 :
Schéma de principe du projet de téléphone pour malentendant TELMA.

MPEG-4 et TEMPOVALSE

Ce principe de codage vidéo par animation de clone virtuel a été, par exemple, l'objet du projet RNRT TEMPOVALSE auquel ont participé le LIS et l'ICP en partenariat avec France Telecom ([Bailly, 2003], voir figure 1.2). L'objet de ce projet était de créer un prototype de terminal portable capturant le mouvement des lèvres à un bout de la ligne puis animant à l'autre bout un visage synthétique dans la norme MPEG-4 ([MPEG, 1997]) qui est un codage audiovisuel basé objet.

Si le problème de bande passante est moins sensible aujourd'hui qu'il ne l'était voici seulement trois ou quatre ans (voir les nombreux systèmes de visioconférence par ADSL commençant à se démocratiser, comme AOL Instant Messenger), le format MPEG-4 demeure un standard dont les diverses fonctionnalités peuvent être utilisées pour mettre en valeur les techniques de segmentation labiale comme par exemple la vidéoconférence virtuelle ou les projets de téléphone pour malentendants comme TELMA (voir figure 1.3).

Ce format de compression considère, en effet, une séquence audiovisuelle comme une composition d'objets multimodaux codés de façon indépendante. MPEG-4-SNHC (Synthetic/Natural Hybrid Coding) dispose de modèles pour représenter les personnes et particulièrement les visages par un jeu de paramètres prédéfinis.

Les paramètres utilisés par MPEG-4 sont :

- les FDP (Facial Definition Parameters) qui contiennent les informations spécifiques à un locuteur, dont les positions d'un jeu de points définissant sa morphologie (les FP, Features Points) et d'autres caractérisant les textures correspondant à l'apparence du visage.
- les FAP (Facial Animation Parameters) permettent d'animer les FP pour suivre les mouvements faciaux correspondant aux différents visèmes ou expressions possibles.
- les FIT (Facial Interpolation Tables) donnent la façon d'interpoler les FAP.

La figure 1.4 présente les points caractéristiques utilisés pour le visage.

Dans le cas d'application type TEMPOVALSE, les FDP sont, soit connus, soit définis sur la première image d'une vidéo, les FAP étant ensuite déterminés en temps réel et transmis. Il est à noter que ce type de définition permettrait d'animer le visage d'une personne avec les mouvements d'une autre, simplement en prenant les FDP de la première et les FAP de la seconde, si FAP et FDP étaient réellement des jeux de paramètres indépendants. Néanmoins, dans [Bailly, 2001], l'auteur a montré que cette hypothèse n'était pas réellement vérifiée car le modèle du visage ne tient pas compte de certains

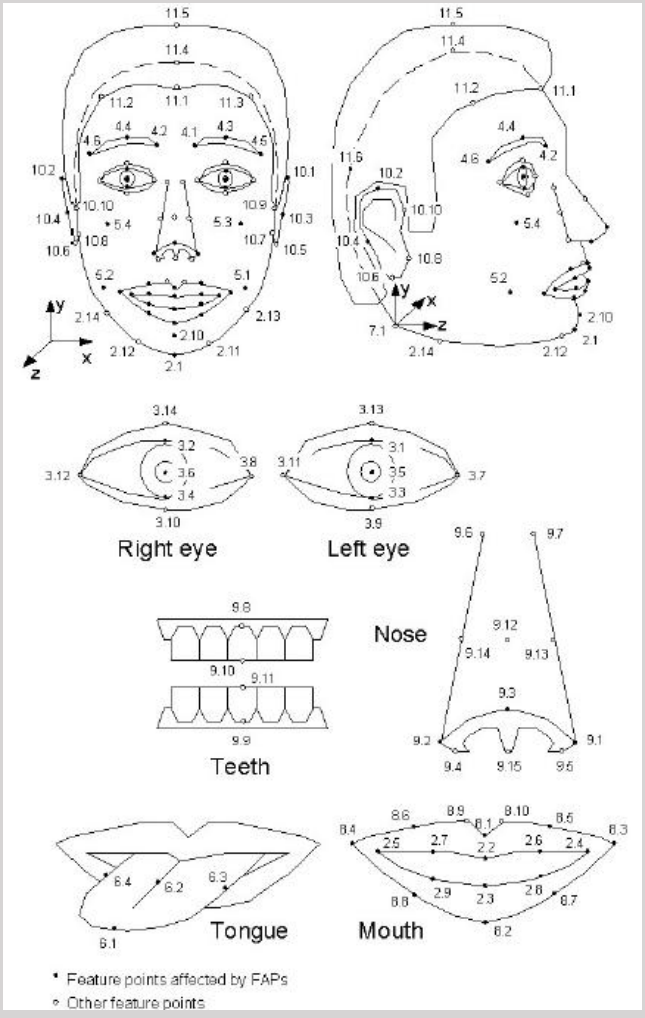


Figure 1.4 : Points de contrôles du MPEG-4 pour le visage.

mouvements articulatoires.

Sous cette réserve, le MPEG-4 peut néanmoins donner lieu à diverses applications dans le domaine du divertissement. Par exemple, les progrès en image de synthèse permettent aujourd'hui de créer des films d'animation mettant en scène des personnages virtuels dont les animations faciales peuvent être obtenues à partir du jeu de véritables acteurs (synchronisation des lèvres, lip-sync)..

Le développement des communautés Internet et des jeux vidéos en ligne, où les utilisateurs sont souvent représentés par l'intermédiaire d'un avatar, est un autre domaine d'application : si apparaître aux yeux des autres sous les traits de personnages fictifs ou réels est déjà possible, bientôt cet avatar pourra être animé et rendre compte en temps réel des émotions des connectés.

Dans [Ostermann, 2004], l'apport que pouvaient amener des clones synthétiques animés par le MPEG-4 dans des interfaces hommes-machines sur des sites marchant a par exemple été démontré.

Autres systèmes de têtes parlantes

Si le format MPEG-4 est un standard, de nombreux systèmes et méthodes permettent de créer et d'animer des personnages virtuels.

En fournissant les déplacements de points de contrôle, le logiciel libre Blender ([Blender]) permet de monter ses propres animations 3D très facilement.

[Ezzat, 2002] présente un système permettant de générer des vidéos synthétiques de locuteurs si réalistes qu'un observateur ne peut les différencier d'une vidéo naturelle. Pour cela, un modèle multidimensionnel déformable est utilisé pour générer des images inédites en déformant les images de la base d'apprentissage, et une modélisation des trajectoires possibles des paramètres du modèle afin que les déformations paraissent naturelles.

Dans [Cosker, 2004], les auteurs ont proposé une méthode utilisant un Modèle Actif d'Apparence (AAM) avec une modélisation non linéaire de la distribution des paramètres du modèle. Un couplage entre audio et apparence du locuteur est ainsi obtenu, permettant de générer des vidéos réalistes d'un visage parlant à partir d'un signal audio. Il est à noter que si l'apparence du modèle n'est entraînée qu'à partir d'un seul locuteur, des sources audio provenant d'individus différents peuvent être utilisées en entrée du système.

1.1.3 Reconnaissance automatique de la parole

La démocratisation progressive des ordinateurs a conduit au développement d'interfaces de plus en plus intuitives entre l'homme et la machine. La souris et les interfaces graphiques font partie de cet effort visant à rendre accessibles et ergonomiques les ressources des ordinateurs personnels. La commande vocale découle de la même volonté et constitue l'une des utilisations les plus évidentes des techniques de reconnaissance automatique de la parole.

N'utilisant d'abord que le canal audio pour des résultats mitigés, les chercheurs de ce domaine ont rapidement tenté d'exploiter l'apport du visuel pour la compréhension de l'oral une fois que celui-ci fut démontré et quantifié (figure 1.1). De plus en plus de

méthodes ont utilisé cette information pour augmenter leurs performances au cours des dix dernières années et plus particulièrement les scores de reconnaissance en milieu bruité. Historiquement, ce sont d'ailleurs les besoins de la reconnaissance de la parole qui ont engendré les premiers travaux sur la segmentation des lèvres.

Si le premier algorithme exploitant l'information visuelle ([Petajan, 1984]) se contentait d'une description très générique des lèvres (hauteur, largeur, surface) et fusionnait les données audiovisuelles de façon séquentielle (le visuel servant uniquement à résoudre les ambiguités de l'audio), les techniques de fusion de données se sont sophistiquées et ont commencé à prendre en compte toute la finesse de description des lèvres que pouvaient fournir les algorithmes de segmentation.

Une vue d'ensemble sur ce domaine de recherche et les différentes approches envisagées ainsi que les types d'architectures de fusion utilisées est proposé dans [Potamianos, 2004]. Deux options se présentent pour cette architecture : la combinaison des vecteurs d'information (fusion précoce, par exemple [Dupont, 2000]), ou la combinaison des probabilités de reconnaissance liées aux modalités (fusion tardive, par exemple [Potamianos, 2004]).

Les modèles de formes et d'apparences actives utilisés dans le cadre de cette thèse ont été mis à contribution à plusieurs reprises : [Faruquie, 2000] présente par exemple un système de reconnaissance de vocabulaire utilisant un Modèle de Forme Actif (ASM, [Cootes, 1995]) où la forme est décrite par des courbes paraboliques. Dans [Dupont, 2000], une variante des ASMs (avec une description de l'apparence prise orthogonalement aux contours au niveau des points de contrôle) est utilisée pour segmenter les lèvres. Les auteurs procèdent ensuite à la reconnaissance automatique de la parole en fusionnant les données audiovisuelles grâce à un Modèle de Markov Cachés (MMC) (voir figure 1.5). En utilisant également un MMC pour la reconnaissance, [Matthews, 2002] a, quant à lui, démontré que les AAMs ([Cootes, 1998]) pouvaient fournir des résultats intéressants en lecture labiale pure (c'est-à-dire sans utiliser le canal audio).

D'autres méthodes ont mis en œuvre des réseaux de neurones de type perceptrons multi-couches pour effectuer la fusion ([Talle, 1997], par exemple).

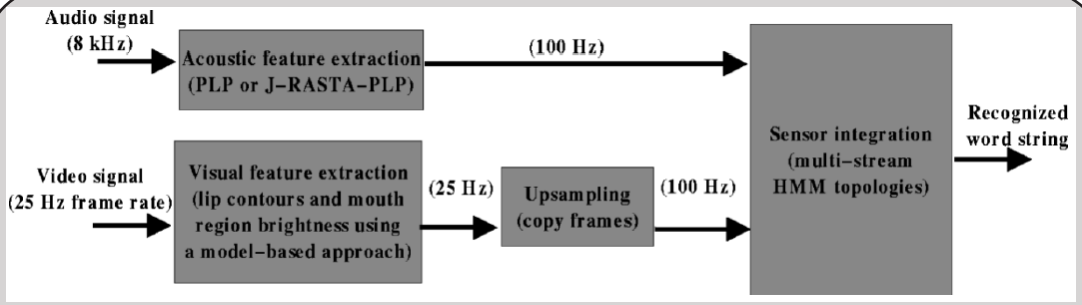


Figure 1.5 :
Architecture d'un module de reconnaissance de parole avec une fusion précoce des données, d'après [Dupont, 2000].
Les informations audio et visuelles pertinentes sont extraites puis fusionnées grâce à un Modèle de Markov Caché.

1.1.4 Reconnaissance d'émotions

Un des autres domaines en vogue des interfaces hommes-machines est la reconnaissance automatique des émotions, l'objectif étant qu'à terme les humeurs d'un utilisateur puissent être prises en compte.

Le Facial Action Coding System ([Ekman, 1978]) est un système de codage de l'activité du visage où l'ensemble des mouvements faciaux possibles sont décomposés en un ensemble de mouvements élémentaires au nombre de 44 et appelés Action Units (AU). Le postulat généralement admis est qu'en identifiant les différentes Action Units sur un visage, on peut classifier leur combinaison comme correspondant à une émotion précise.

La plupart des travaux actuels tentent ainsi d'identifier un nombre encore relativement limité d'émotions génériques : joie, tristesse, surprise, peur, dégoût, colère (l'expression « neutre » étant parfois considérée comme une émotion à part entière) en détectant des indices visuels sur le visage.

Dans ce cadre, la bouche est naturellement un indice très pertinent pour ce type de travaux, mais les algorithmes détectent également généralement la forme des yeux, des sourcils, le nez ou le sillon nasolabial.

[Pantic, 2000] fournit un état de l'art des techniques de reconnaissance d'expression.

Dans [Pantic, 2000+], par exemple, une segmentation de divers indices visuels, dont la bouche, par une méthode contour actif ([Kass, 1987]) est utilisée pour déterminer l'expression grâce à un réseau de neurones effectuant une classification floue. Dans [Tian, 2000], les auteurs s'étaient intéressés à l'identification d'AUs situés sur le bas de visage en utilisant un modèle analytique pour décrire les lèvres et le sillon nasolabial et un réseau de neurones à trois couches pour la partie reconnaissance. Plus récemment, [Pantic, 2006] a utilisé le filtrage particulaire pour suivre un jeu de 15 points faciaux donnant accès au comportement temporel de 27 des AU grâce à un système de règles statiques et dynamiques. [Zhang, 2005] propose d'utiliser les Réseaux Bayésiens pour déterminer l'état émotionnel en suivant un jeu de 26 points caractéristiques (dont 8 sur le contour des lèvres) grâce à

Figure 1.6 :
Système de reconnaissance et de synthèse d'expression faciale, d'après [Abboud, 2004].
De gauche à droite : visage inconnu joyeux, adaptation du modèle, annulation de l'expression et synthèse d'un visage neutre, génération de six expressions faciales synthétiques (colère, dégoût, peur, joie, surprise, tristesse).

une description locale des points d'intérêt par des filtres de Gabor ([Daugman, 1980]) et un Filtrage de Kalman.

Dans [Aleksic, 2005], il a par ailleurs été montré que les FAPs du standard MPEG-4 pouvaient être utilisés pour identifier l'état émotionnel d'une personne en utilisant une approche par MMC.

Les AAMs ont, quant à eux, été utilisés dans [Abboud, 2004] pour effectuer de la reconnaissance ainsi que de la synthèse d'expression. A partir du jeu de paramètres contrôlant le modèle actif, une Analyse Discriminante Linéaire est effectuée pour déterminer à quelle classe d'expression correspond le visage traité. Une fois l'expression connue, l'apparence peut alors être ramenée à l'expression neutre pour ensuite synthétiser n'importe quelle des 6 expressions possibles (voir figure 1.6).

1.2 Etat de l'art de l'analyse labiale

Dans cette partie nous allons procéder à un état de l'art du domaine de la segmentation labiale en regroupant les méthodes en trois catégories selon leurs approches:

- approche « apparence » ou « pixel » : les lèvres sont segmentées en exploitant les valeurs des pixels dans les espaces couleurs
- approche « forme » : méthode détectant les contours des lèvres en utilisant des modèles plus ou moins complexes de la forme de la bouche
- approche « forme et apparence » : méthode utilisant des modèles pour décrire à la fois la forme et l'apparence

Il est à noter que ces approches ne répondent pas forcément aux mêmes objectifs de précision et finesse et ne sont pas nécessairement mutuellement exclusives, ainsi une méthode de type «pixel» peut par exemple être utilisée comme prétraitement d'une méthode utilisant un modèle de lèvre.

Cet état de l'art détaillera plus particulièrement les approches utilisant les Modèles Actifs que ce soit de Forme (ASM) ou d'Apparence (AAM).

1.2.1 Méthode avec une approche pixel

Cette famille de méthodes regroupe toutes les approches exploitant la distribution des couleurs présentes dans une image, sans considération de forme ou de contour. Elle prennent donc comme postulat que les lèvres correspondent à un groupe de pixels de caractéristiques homogènes dans un espace couleur donné et vont tenter de segmenter l'image en des régions « lèvres » et « autre ».

L'espace couleur utilisé le plus fréquemment employé est le classique RVB (Rouge Vert Bleu), mais le YCbCr (où Y est la luminance, et Cb et Cr sont des composantes chromatique), le TLS (Teinte, Saturation, Luminance, qui est le système se rapprochant le plus de la perception humaine) et les espaces perceptifs (Luv, Lab) font également partie des espaces utilisés régulièrement pour cette tâche. Voir par exemple [Ford, 1998] pour une présentation des divers espaces couleurs.

Les techniques les plus élémentaires de cette famille reviennent ainsi à faire des

seuillages sur une grandeur colorimétrique jugée pertinente. Ce principe est par exemple utilisé dans [Petajan, 1984], article fondateur déjà évoqué précédemment, où l'on procédait à un simple seuillage de l'image de luminance et encore récemment dans [Lyons, 2003] où un deuxième seuillage est effectué sur la composante R du RVB.

Pour s'affranchir de la dépendance à l'éclairage des simples niveaux de gris ou du RVB, de nombreux auteurs ont proposé d'utiliser d'autres espaces couleurs, où la séparation entre les lèvres et le reste du visage serait facilitée grâce à l'utilisation de composantes chromatiques.

Dans [Zhang, 2000], les auteurs exploitent par exemple la Teinte, tandis que [Hsu, 2002] introduit une grandeur synthétique Cr/Cb-Cr2.

De nombreuses méthodes de classification statistique ont également été utilisées afin d'effectuer cette segmentation de façon plus sophistiquée que par un simple seuillage.

Par exemple, [Liew, 1999] traite de l'utilisation des techniques d'agrégation floues, où les pixels sont classés à partir d'une carte d'appartenance aux lèvres dans l'espace Luv et Lab.

Les champs de Markov aléatoires ont été utilisés dans [Liévin, 2004] afin d'obtenir une segmentation hiérarchique incluant des paramètres colorimétriques mais aussi de mouvement, avec la particularité que le voisinage spatiotemporel des pixels est pris en compte lors de la classification.

Enfin, diverses techniques nécessitant au préalable un étiquetage manuel d'une

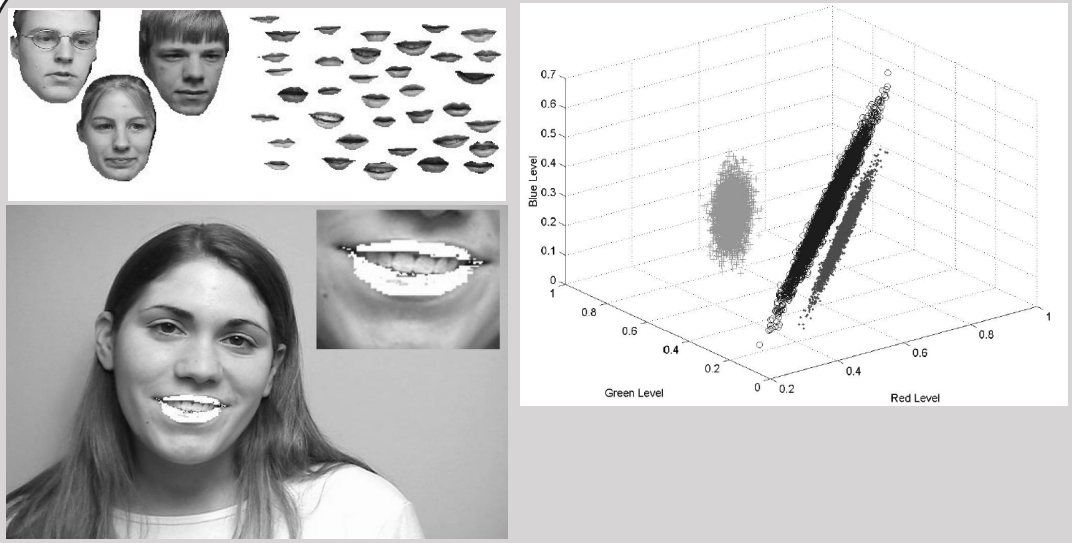


Figure 1.7 :
Système de détection de lèvre par modélisation gaussienne des distributions des valeurs de peau, lèvres et fond dans l'espace RVB, d'après [Patterson, 2002].
De gauche à droite et de haut en bas : base d'apprentissage manuellement segmentée, distribution du fond (+), du visage (o) et des lèvres (.) et exemple de segmentation automatique.

base d'apprentissage afin d'obtenir des connaissances colorimétriques de référence ont également été développées.

L'Analyse Discriminante Linéaire a par exemple été utilisée dans [Nefian, 2002] afin d'obtenir une combinaison des composantes de l'espace effectuant la meilleure discrimination possible entre les lèvres et le reste du visage. Dans [Patterson, 2002], les auteurs se servent des données d'apprentissage pour obtenir une estimation gaussiennes des distributions statistiques pour les classes de peau, de lèvre et de non-visage dans l'espace RVB, la classification étant ensuite effectuée par un classification bayesienne (figure 1.7).

L'absence de contrainte sur la forme et l'orientation «région» de la recherche fait que ce type de méthode ne donne en général que des formes approximatives des lèvres et est particulièrement sensible au bruit. Néanmoins, les résultats donnent en général des informations élémentaires sur les lèvres (largeur, hauteur) qui peuvent être suffisantes pour certaines applications. Les parties suivantes vont présenter des approches orientées «contours» visant à obtenir des bouches détectées plus réalistes et plus fines, la partie 1.2.3 présentant quelques exemples de méthodes orientées pixels rendues plus robustes par l'apport d'informations sur la forme. Dans le cadre de ce travail de thèse, une méthode appartenant à la dernière sous-famille évoquée a été développée afin d'être utilisée comme pré-traitement des modèles actifs.

1.2.2 Méthodes avec une approche « forme »

Cette famille de méthode vise à effectuer la segmentation des lèvres par une approche « contour ». Les valeurs des pixels sont utilisées pour faire converger la recherche sur les contours recherchés mais de façon indirecte, comme par exemple par l'intermédiaire des images gradients.

On peut alors séparer les méthodes selon qu'un modèle est utilisé ou non pour décrire les lèvres et, le cas échéant, le degré de complexité et la nature du modèle utilisé. On obtient alors trois catégories :

- les méthodes sans modèle spécifique de lèvres
- les méthodes utilisant un modèle analytique
- les méthodes utilisant un modèle statistique.

1.2.2.1 Méthodes sans modèle de lèvres

Les exemples les plus communs de ce type d'approche sont les méthodes faisant appel aux contours actifs ou « snakes ». Cette méthode populaire a été introduite par [Kass, 1987] et désigne des courbes paramétriques flexibles ayant la capacité de se déformer de façon à converger sur les contours d'un objet quelconque.

Cette convergence est effectuée en minimisant une fonction de coût possédant deux termes :

- l'énergie interne qui permet de régulariser l'élasticité et la courbure du contour (deux paramètres à régler empiriquement permettant de contrôler le degré de flexibilité accordé au snake).

- l'énergie externe qui va permettre de prendre en compte les informations de l'image sous la forme d'une force extérieure (définie par exemple en fonction de la proximité du snakes avec les zones de fort gradient, caractéristiques des contours) ainsi que d'autres contraintes éventuelles (il est possible par exemple de définir des forces «ressorts» liant le snakes à des points fixes ou des forces « ballons » qui incitent le snake à se contracter ou à se dilater en l'absence de forces extérieures marquantes).

L'optimisation de la fonction de coût s'effectue très simplement grâce à un calcul matriciel itératif. Si le snake est initialisé à proximité de l'objet cible dont les contours sont suffisamment marqués, la segmentation pourra être de bonne qualité.

Les snakes pouvant s'adapter potentiellement à tous types d'objets, les appliquer aux contours des lèvres a été envisagé à plusieurs reprises mais a rencontré quelques écueils. Le problème de l'initialisation est en effet très pointu et limitant quand il s'agit d'envisager des applications totalement automatiques et pouvant s'appliquer à tous types de prises de vue, sans contrôle des conditions. En outre, mis à part la contrainte des deux paramètres d'élasticité et de courbure (dont le réglage optimal pour toute situation peut se révéler ardu, voir impossible, à trouver), le snake a une forme totalement libre et peut donc converger vers une forme qui ne ressemblera parfois nullement à des lèvres réalistes.

Un exemple parmi d'autres d'applications des contours actifs à la zone labiale est donné dans [Delmas, 2002]. Néanmoins si les résultats se révèlent très satisfaisants si les conditions sont bonnes (éclairage contrôlé, bon contraste entre la couleur de la peau et celle des lèvres), les performances décroissent lorsque ces conditions favorables ne sont plus réunies, la phase d'initialisation du snake manquant de robustesse ce qui nécessite de complexifier la méthode en procédant en plusieurs phases de convergence (voir figure 1.8).

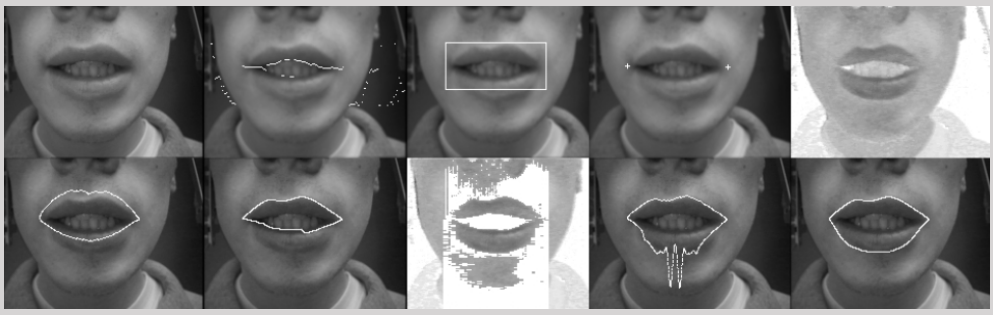


Figure 1.8 :
Exemple de détection du contour extérieur des lèvres par une méthode de contours actifs incluant un repositionnement du snake après mauvaise initialisation et convergence par test d'homogénéité de régions, d'après [Delmas, 2000].
De gauche à droite et de haut en bas : Plan Luminance initial, minima verticaux du plan précédent, région d'intérêt correspondante, coins de la bouche, plan teinte, snake initial, mauvaise convergence du snake, statistique d'appartenance à la zone lèvre en découlant, initialisation du snake correspondante, convergence finale.

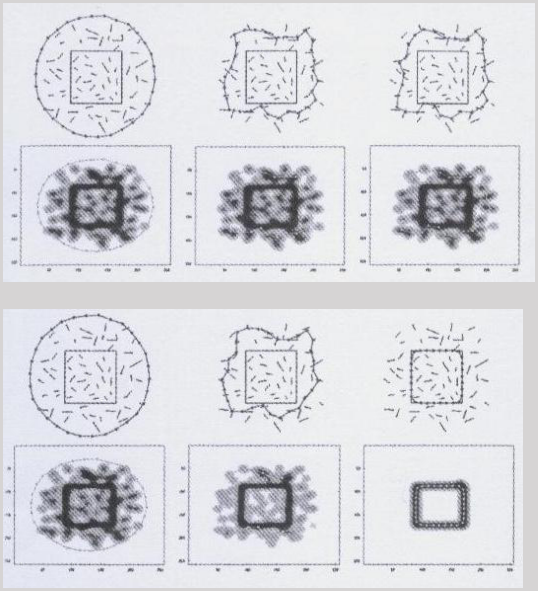


Figure 1.9 :
Snake adaptatif dont le champ de potentiel évolue au cours de la convergence, d'après [Nascimento, 2005].
En haut, convergence d'un snake non adaptatif avec le champ de potentiel correspondant. En bas convergence d'un snake adaptatif avec évolution du champ de potentiel au fur et à mesure de la convergence vers un résultat correct.

Plus récemment, dans [Nascimento, 2005], il a été proposé d'utiliser des snakes «adaptatifs» utilisant un algorithme d'Estimation/Maximisation (EM, [McLachlan, 1997]). Les points de forts gradients de l'image (correspondant aux contours) sont séparés en plusieurs segments qui vont être classifiés comme étant valides (sur l'objet à détecter) ou non valides (sur d'autres d'objets ou sur des contours intérieurs de l'objet recherché) avec un certain degré de confiance qui est réévalué par l'algorithme EM à chaque itération de la convergence du snake. Cela revient à rendre le champ de forces extérieures guidant la convergence du snake dynamique et adaptatif au fur et à mesure de la convergence permettant en quelque sorte de débruiter l'image (la figure 1.9 illustre ce principe dans un cas simple).

Cette famille de méthodes présentent des résultats satisfaisants dans de bonnes conditions mais, de façon similaire aux méthodes du 1.2.1, l'absence de réelle contrainte sur la forme finale de l'objet détecté fait que l'on n'a pas de réelle certitude sur la pertinence du résultat et l'on a une grande sensibilité au bruit. Les deux parties suivantes vont intégrer le principe d'un modèle de la forme de la bouche.

1.2.2.2 Méthodes avec des modèles de lèvres analytiques

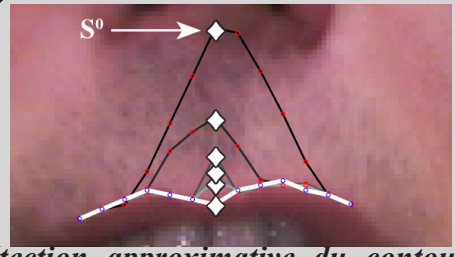
Dans l'objectif d'avoir un contour détecté plus conforme à la réalité par rapport aux méthodes n'utilisant que l'apparence ou n'ayant pas de modèle de la forme des lèvres, de très nombreux auteurs ont proposé d'utiliser des modèles paramétriques ou patrons déformables (deformable templates).

Les lèvres seront alors décrites par un certain nombre de points de contrôle pertinents et de courbes qui constituent une forme prototype pouvant se déformer en jouant sur un jeu de paramètres de contrôle. L'une des difficultés de ces approches est de trouver le

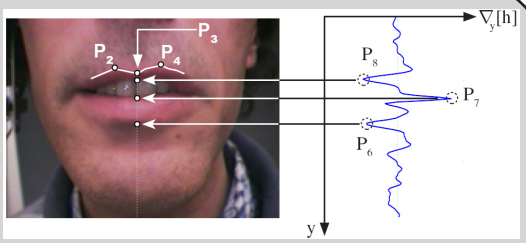
bon dosage de flexibilité : un modèle peu flexible donnera toujours une forme de lèvre plausible mais échouera parfois à segmenter des formes de bouches ou des configurations moins habituelles alors qu'un modèle trop flexible donnera des formes non réalistes dans certains cas. Une fois le modèle analytique déterminé, il convient encore de déterminer des critères pour paramétrer correctement le modèle sur une image inconnue.

A titre d'exemple, dans [Horbet, 1995], les auteurs ont décidé d'améliorer l'approche par contours actifs en rajoutant une force intérieure baptisée « force template ». A chaque itération de la convergence du snake, la force template ramène le contour actif sur une forme admissible de lèvre. L'utilisation du template fournit en quelque sorte une connaissance a priori des formes possibles et contraint le snake à des déformations encadrées.

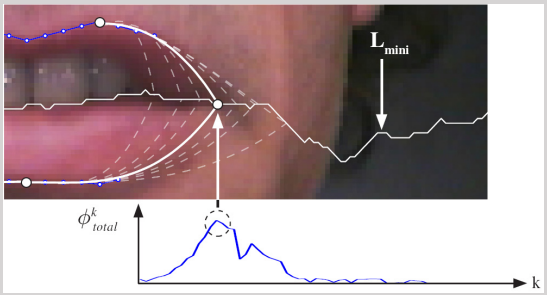
Dans [Hennecke, 2000], un patron de lèvres (constitué de trois quartiques pour le contour extérieur et de deux paraboles pour le contour intérieur et contrôlé par un jeu de 12 paramètres) est utilisé. Les valeurs optimales de ces paramètres sont obtenues en minimisant une fonction de coût utilisant le gradient vertical de l'image comme champ de



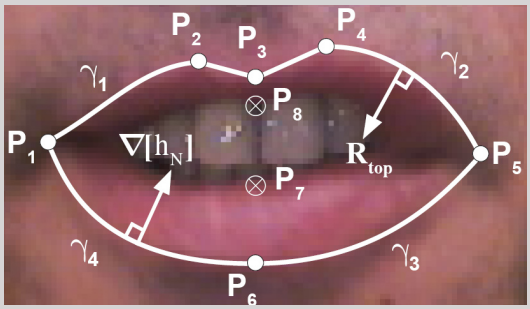
Détection approximative du contour supérieur par un algorithme de snake "bondissant" qui est initialisé en un germe S^0 au-dessus des lèvres et qui tombe jusqu'à rencontrer le contour.



A partir de la convergence du snake, détection de points de contrôle en considérant le profil du gradient de teinte h sur la verticale.



Recherche des commissures des lèvres sur la ligne des minima de luminance et optimisation des paramètres de cubiques en maximant un flux de gradient à travers la courbe.



Modèle analytique des lèvres avec 6 points de contrôle, 4 cubiques et 2 segments de droites pour l'arc de cupidon.

Figure 1.10 :
Exemple de détection du contour extérieur des lèvres par une méthode de modèle analytique déformable constitué de 4 cubiques, d'après [Eveno, 2004].

potentiel.

Dans un souci d'avoir un modèle suffisamment flexible afin de pouvoir modéliser n'importe quelle configuration de lèvre, il a été proposé dans [Eveno, 2004] d'utiliser quatre courbes paramétriques cubiques pour décrire le contour extérieur des lèvres en imposant conditions et limites aux dérivées des courbes au niveau des points saillants. Après une initialisation où un snake simplifié (sans force intérieure) donne une première estimation du contour supérieur et la détection d'un point sur le contour inférieur, les paramètres des cubiques sont optimisés en maximisant le flux de vecteurs gradients à travers les différentes courbes (voir figure 1.10).

Les méthodes de cette famille donnent parfois d'excellents résultats, en fonction de la pertinence de la façon dont les auteurs auront paramétré la flexibilité de leur bouche déformable. L'étape de création et de calibration d'un modèle analytique étant néanmoins très longue, des auteurs ont cherché à obtenir des modèles directement à partir d'un ensemble de données réelles par une approche statistique.

1.2.2.3 Méthode utilisant un modèle statistique de la forme

Cette catégorie de méthode correspond presque exclusivement aux Modèles Actifs de Forme (ASM, [Cootes, 1995]). Les ASMs correspondent en fait à l'application des Modèles de Distribution de Points (Point Distribution Model, PDM, [Cootes, 1992]) pour segmenter un objet sur une image.

Principe général

Les PDMs sont construits à partir d'exemples d'entraînements : les contours de l'objet que l'on veut modéliser seront représentés par un nombre N de points qui doivent être étiquetés manuellement sur un assez grand nombre M d'images.

On va alors disposer de M vecteurs $\boldsymbol{x}_i$ contenant les coordonnées des points (dans un espace à deux ou trois dimensions selon les cas) qui, après une indispensable normalisation (par exemple grâce à la transformation procrustéenne généralisée), constitueront la base d'apprentissage du PDM.

On calcule alors le vecteur moyen :

$$\bar{\boldsymbol{x}} = \frac{\mathbf{1}}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \boldsymbol{x}_i \qquad \text{(eq. 1.1)}$$

puis l'on procède à l'Analyse en Composante Principale (ACP) de la matrice de covariance centrée. Toute forme pourra alors être approximée grâce à l'équation :

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \mathbf{P}_s \boldsymbol{b}_s \qquad \text{(eq. 1.2)}$$

qui correspond à la somme pondérée des n vecteurs propres $\boldsymbol{p}_i$ les plus significatifs (qui correspondent à une portion choisie de la variance totale) qui sont rangés dans la matrice $\mathbf{P}_S = [\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n]$ avec $\boldsymbol{b}_S = [b_1, \ldots, b_n]$ un vecteur contenant les poids affectés à chaque

Figure 1.11 :
Exemple d'image annotée et premiers modes de variations de l'ASM (selon [Cootes, 2004]).

mode propre. La figure 1.11 montre un exemple d'image annotée et l'effet des premiers modes propres tels que vus dans [Cootes, 2004] pour un modèle de visage. On notera que l'on peut avoir des problèmes à interpréter les modes propres. L'ASM sera ainsi la méthode permettant d'adapter le PDM et de segmenter un objet sur une image inconnue en déterminant le vecteur de paramètres $\boldsymbol{b_s}$ optimal.

Le modèle associé étant linéaire, on considère que les valeurs b_i ont une distribution gaussiennes et varient entre $-3\sigma_i$ et $+3\sigma_i$ où σ_i est l'écart type associé au vecteur propre p_i.

La figure 1.12 présente un exemple de convergence pour un PDM de visage exploitant une fonction de coût basée sur une maximisation d'un flux de gradients à travers une courbe. La méthode originellement utilisée dans [Cootes, 1995] consistait à déplacer individuellement les points du contour de façon à les rapprocher des zones de forts gradients correspondant aux contours puis à déterminer le vecteur $\boldsymbol{b_s}$ permettant de rendre le mieux compte des modifications.

Par opposition aux modèles paramétriques qui sont bâtis de façon empirique par des chercheurs experts, les modèles statistiques ne nécessitent donc pas de réflexion sur la paramétrisation de patron ou de dosage de flexibilité. Le modèle sera en effet naturellement capable de se déformer de façon à reproduire toute forme présente dans la base d'apprentissage, mais la limite en est justement la phase d'apprentissage. Le temps économisé à paramétrer son patron déformable est en partie perdu lors de l'étiquetage manuel de la base d'apprentissage ; en outre une configuration absente de l'apprentissage sera probablement impossible à segmenter ultérieurement sur une image inconnue, le modèle manquant de flexibilité en dehors de ce qu'il a appris à faire.

Néanmoins, si la base d'apprentissage est suffisamment conséquente et l'étiquetage

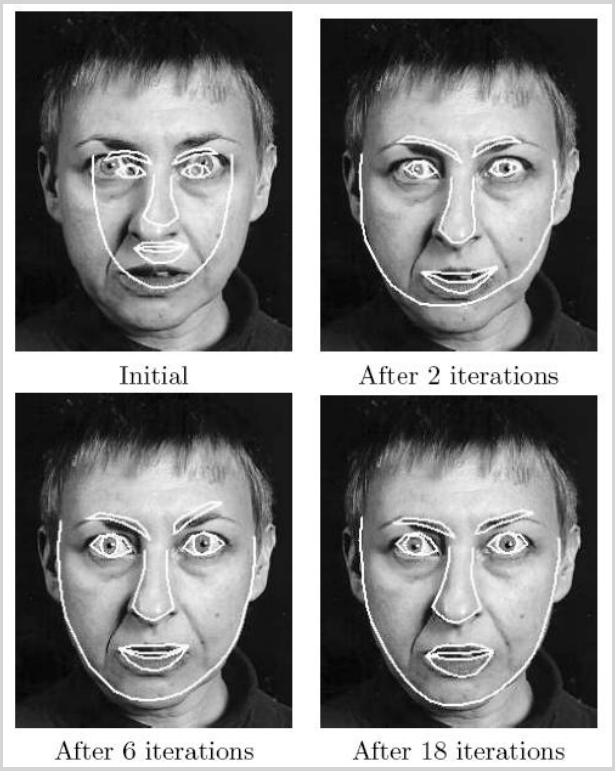


Figure 1.12 : Exemple de convergence d'un ASM dans le cas d'un visage selon [Cootes, 2004].

soigneusement fait, les ASMs se révèlent performants et robustes et, s'ils ont d'abord été utilisés dans le domaine biomédical, leur champ d'application s'est étendu à de nombreux domaines, dont l'application labiale.

La méthode ayant reçu un très bon accueil de la communauté, de nombreux travaux ultérieurs ont complexifié et enrichi le principe de base. Plusieurs articles se sont par exemple intéressés à corriger les éventuels problèmes dus à la linéarité du modèle venant de l'ACP.

Distribution non-linéaire des paramètres

Les ASMs de certains objets présentent des distributions pour le vecteur $\boldsymbol{b_s}$ qui ne correspondant clairement pas à une caractéristique gaussienne et chercher les valeurs b_i sans contraintes entre $-3\sigma_i$ et $+3\sigma_i$ conduit à des formes impossibles. La figure 1.13 présente l'exemple de la base d'apprentissage d'un objet présentant un mouvement de rotation et la distribution des modes du PDM (selon [Cootes, 1997]).

Plusieurs approches ont été envisagées pour corriger ce problème quand il se posait (principalement pour des objets présentant des articulations, des coudes ou des rotations) en rajoutant des éléments non linéaires lors de la construction du PDM.

[Sozou, 1994] a d'abord proposé d'utiliser une régression polynomiale afin de permettre aux points de varier selon des trajectoires polynomiales et non linéaires lorsque le vecteur de paramètre $\boldsymbol{b_s}$ est modifié. Peu de temps après, dans [Sozou, 1997], un perceptron à

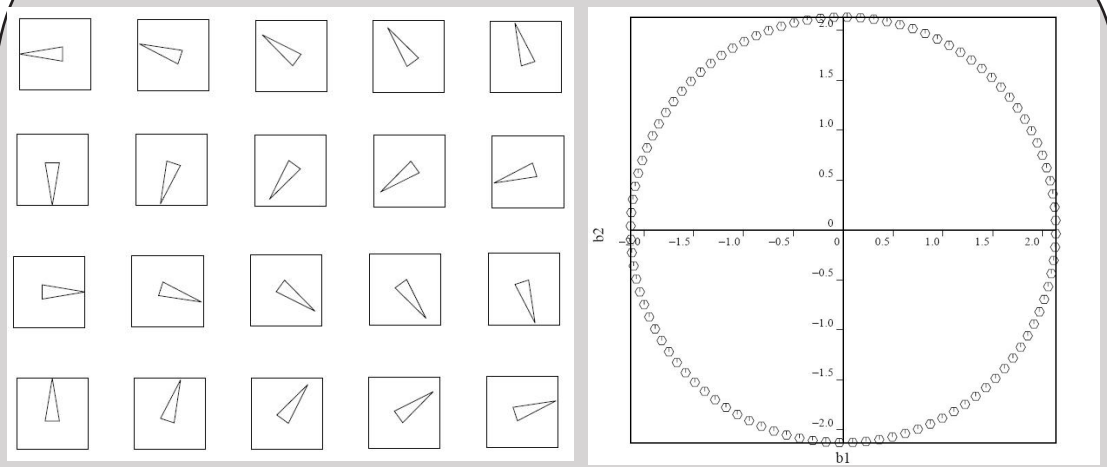


Figure 1.13 :
Cas de distribution non linéaire des modes propres, selon [Cootes, 1997].
A gauche : base d'apprentissage d'un objet présentant des rotations.
A droite : distribution des valeurs prises par les poids affectés aux deux modes propres de l'ASM sur la base d'apprentissage. Comme on constate que cette distribution est circulaire, permettre aux modes propres de varier linéairement et indépendamment l'un de l'autre lors de l'optimisation du vecteur de paramètre b_S conduirait fréquemment à des formes non réalistes. Dans un tel cas de figure, le recours à des modèles non-linéaires est nécessaire.

plusieurs couches a été utilisé pour traiter les objets présentant des rotations qui demeuraient mal traités par la méthode polynomiale. Dans [Cootes, 1997], les auteurs ont également proposé une méthode plus simple en décrivant la distribution du vecteur $\boldsymbol{b_S}$ par un mélange de gaussiennes. Les configurations du vecteur $\boldsymbol{b_S}$ obtenues par la convergence d'un ASM seront alors modifiées pour être ramenées sur la configuration plausible la plus proche.

Si ces approches non linéaires des PDMs sont très intéressantes pour certaines études, leur relative complexité ne se justifie pas dans le cas de l'analyse labiale où la forme des lèvres ne présente pas de comportement non linéaire (absence de rotation importante du visage).

Raffinements des ASMs

Dans [Duta, 1997], il a été proposé une amélioration des ASMs permettant de gérer les problèmes d'occultations partielles de l'objet à détecter en utilisant les connaissances a priori sur la forme pour replacer les points dans les zones cachées. Dans [Romdhani, 1999], une forme non linéaire d'ASM a été mise en oeuvre en utilisant l'ACP à noyau ([Scholkopf, 1998]) afin de pouvoir segmenter le visage humain sous différents angles de

vue sans pour autant avoir recours à un modèle 3D du visage.

Notre travail se plaçant dans l'hypothèse que le visage traité est vu de face et sans occultation, ces possibilités intéressantes n'ont pas été approfondies.

Enfin, dans [Zhang, 2003], dans le cadre d'un PDM appliqué au visage humain entier, il a été proposé une version utilisant des contraintes sur les contours du visage (joues, arc mandibulaire) et une description des points de contrôle du PDM (yeux, nez et bouche) dans une base de filtres de Gabor ([Daugman, 1980]). Les points présents sur les contours du visage doivent minimiser un critère faisant intervenir un modèle local de texture qui sera minimisé si les points sont placés sur des zones de forts gradients. Les points de contrôle du PDM sont, quant à eux, décrits par un jeu de 40 coefficients correspondant aux réponses des filtres de Gabor pour différents angles et échelles. L'utilisation de fonctions de similarités permet de définir à chaque itération des points présentant de meilleures réponses aux filtres. Le PDM intervient enfin pour régulariser à chaque étape de la convergence les déplacements de points sur des formes plausibles.

Enfin, la principale amélioration apportée aux ASMs peut être considérée comme la prise en compte de l'apparence et des valeurs de niveaux de gris, ce qui a conduit aux Modèles Actifs d'Apparence présentés au 1.2.3.

1.2.3 Méthodes avec une approche combinant forme et apparence

Cette famille de méthode comprend tout d'abord un certain nombre de techniques présentant des caractéristiques les plaçant à la frontière entre les méthodes présentées au 1.2.1 et au 1.2.2.

Dans [Liew, 2000], les auteurs ont proposé d'utiliser une carte de probabilité résultant d'une classification floue des pixels en deux zones 'lèvre' et 'non lèvre' pour optimiser les paramètres d'un modèle déformable à trois paraboles décrivant le contour extérieur des lèvres.

Dans [Tian, 2000], une méthode est présentée où un modèle de forme analytique à plusieurs états (ouvert, relativement fermé, étroitement fermée) est utilisé en parallèle à une description de la distribution colorimétrique des lèvres par une mixture de gaussienne. Un algorithme de suivi de mouvement inspiré par la méthode Lucas-Kanade ([Lucas, 1981]) combinant les informations de forme et de couleurs permet alors d'effectuer la segmentation.

De nombreux autres exemples pourraient être donnés, mais comme les principes ont déjà été présentés dans les parties précédentes, cette partie va principalement concerner les modèles statistiques.

Première modélisation statistique de l'apparence

Très rapidement après la création des ASMs, la volonté de modéliser non seulement la forme mais également l'apparence a en effet donné naissance à des méthodes utilisant une description des niveaux de gris présents sur l'image.

Dès [Cootes, 1993], l'idée d'un modèle statistique de profil de niveau de gris autour des

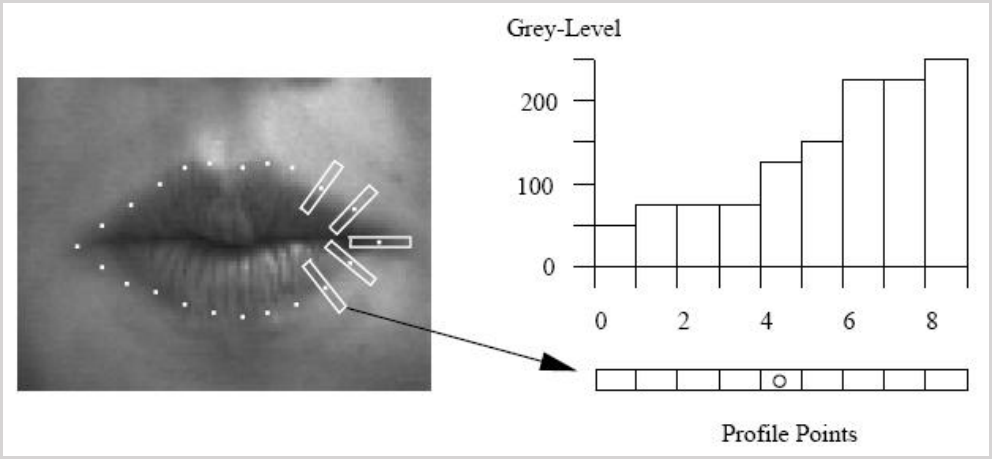


Figure 1.14 :
Profil de niveaux de gris utilisé par [Luettin, 1996].

points du PDM a été introduite, principe repris plus tard dans [Luettin, 1996] et illustré par la figure 1.14. Une ACP est alors effectuée sur une base de vecteurs contenant les valeurs prises par les pixels situés sur des segments normaux aux contours recherchés ce qui permet d'obtenir une équation similaire à l'équation 1.2 :

$$\boldsymbol{pg} = \overline{\boldsymbol{pg}} + \mathbf{P}_{\mathbf{pg}}\boldsymbol{b}_{pg} \qquad \text{(eq. 1.3)}$$

où $\boldsymbol{pg}$ est le vecteur contenant les profils de niveaux de gris, $\boldsymbol{b}_{pg}$ le vecteur de poids appliqués aux modes propres contenus dans $\mathbf{P}_{\mathbf{pg}}$.

Lors de la convergence de l'ASM, si l'on considère que $\boldsymbol{pg}$ correspond au profil observé pour une configuration du vecteur $\boldsymbol{b}_{\boldsymbol{s}}$ contrôlant la forme, on appelle $\boldsymbol{pg}_{\boldsymbol{bf}}$ (bf pour best fit) la meilleure approximation du profil par le modèle qui est obtenu ainsi :

$$\boldsymbol{b}_{pg\_bf} = \mathbf{P}_{\mathbf{pg}}{}^{T}\left(\boldsymbol{pg} - \overline{\boldsymbol{pg}}\right) \Rightarrow \boldsymbol{pg}_{bf} = \overline{\boldsymbol{pg}} + \mathbf{P}_{\mathbf{pg}}\boldsymbol{b}_{pgbf} \qquad \text{(eq. 1.4)}$$

La convergence de l'ASM s'obtient alors en minimisant le carré de l'erreur d'estimation des profils E_p :

$$E_p{}^2 = \left(\boldsymbol{pg}_{bf} - \boldsymbol{pg}\right)^T\left(\boldsymbol{pg}_{bf} - \boldsymbol{pg}\right) = \left(\boldsymbol{pg} - \overline{\boldsymbol{pg}}\right)^T\left(\boldsymbol{pg} - \overline{\boldsymbol{pg}}\right) - \boldsymbol{b}_{pgbf}{}^T\boldsymbol{b}_{pgbf} \qquad \text{(eq. 1.5)}$$

ce qui revient donc à réduire l'écart entre les profils observés et leurs meilleures approximations possible.

Après l'apparition d'un modèle statistique d'apparence, utilisé pour le suivi et le codage de visage ([Lanitis, 1994]), où les niveaux de gris d'un visage étaient modélisés par ACP

dans leur ensemble, les Modèles Actifs d'Apparence (AAM) furent réellement introduits dans [Cootes, 1998].

Principe des Modèles Actifs d'Apparence

L'idée force de [Cootes, 1998] est de lier deux modèles statistiques actifs (construits quasi indépendamment l'un de l'autre par ACP) décrivant la forme et les niveaux de gris par un troisième modèle statistique construit par une ACP effectuée sur les valeurs des vecteurs de poids.

Comme pour un ASM, la construction d'un AAM nécessite de disposer d'une base de M images d'apprentissages manuellement annotées.

La construction du modèle de forme est similaire à l'ASM et il sera contrôlé par l'équation 1.2. En revanche pour construire le modèle de niveaux de gris, les visages sont préalablement déformés par triangulation de façon à ce que les points du maillage se retrouvent alignés sur la forme moyenne.

Les valeurs de niveaux de gris de la région recouverte par la forme moyenne sont alors prélevées sur ces images normalisées par rapport à la forme et placées dans des vecteurs $\boldsymbol{g}$. Ces vecteurs peuvent alors être normalisés pour diminuer l'influence de l'illumination en affectant un offset et coefficient d'échelle, mais seront toujours notés $\boldsymbol{g}$ par la suite.

On procède alors à une ACP pour parvenir à une équation similaire à la 1.2 :

$$\boldsymbol{g} = \bar{\boldsymbol{g}} + \mathbf{P}_g \boldsymbol{b}_g \qquad \text{(eq. 1.6)}$$

Toute forme $\boldsymbol{x}$ et apparence $\boldsymbol{g}$ peuvent alors être représentés par les jeux de coefficients les approximant le mieux $\boldsymbol{b}_s$ et $\boldsymbol{b}_g$. $\mathbf{W}_s$ étant une matrice diagonale de normalisation des unités, on procède alors à une nouvelle ACP sur les vecteurs b définis comme suit :

$$\boldsymbol{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{P}_s^T \left(\boldsymbol{x} - \bar{\boldsymbol{x}} \right) \\ \mathbf{P}_g^T \left(\boldsymbol{g} - \bar{\boldsymbol{g}} \right) \end{pmatrix} \qquad \text{(eq. 1.7)}$$

et l'on obtient finalement l'équation contrôlant l'AAM :

$$\boldsymbol{b} = \mathbf{P}_c \boldsymbol{c} \qquad \text{(eq. 1.8)}$$

où $\mathbf{P}_c$ contient les vecteurs propres retenus (le vecteur moyen étant nul, $\boldsymbol{b}$ ayant une valeur nulle du fait que les poids $\boldsymbol{b}_s$ et $\boldsymbol{b}_g$ ont des distributions gaussiennes centrées sur zéro).

La figure 1.15 présente les modes de variation de forme, de niveaux de gris et enfin d'apparence. Il est à noter que la modélisation linéaire de l'apparence entraîne certaines limites dues à certains phénomènes non-linéaires sur le visage (comme l'apparition de rides ou de plis sur la peau) et sur la zone labiale en particulier (apparition/disparition des dents, par exemple).

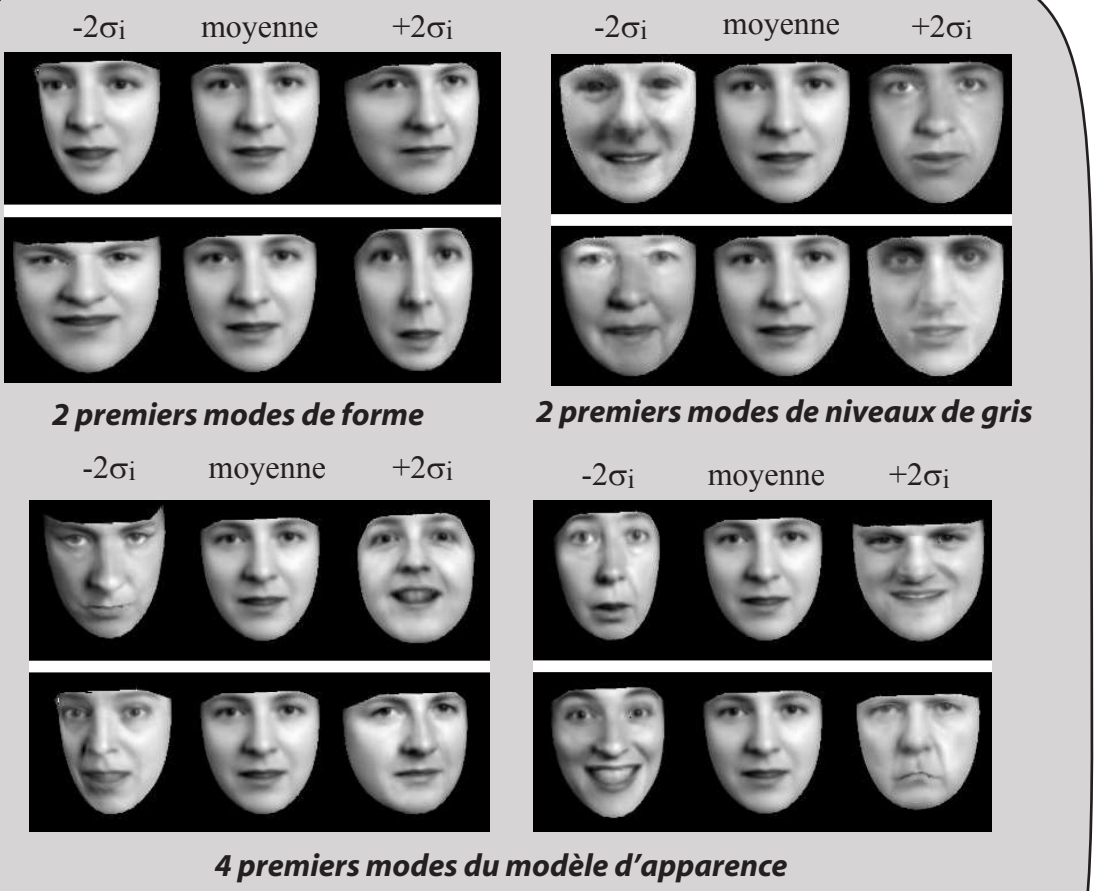


Figure 1.15
Premiers modes de variations pour les différents modèles d'un AAM, [Cootes, 1998].

Pour segmenter un visage inconnu, il faut donc trouver le vecteur optimal de paramètres $\boldsymbol{c}$ ainsi que des paramètres de position $\boldsymbol{t}$ et d'échelle $\boldsymbol{u}$. Dans la suite on considérera le vecteur de paramètre global $\boldsymbol{p}$ tel que : $\boldsymbol{p}^T=[\boldsymbol{c}^T ; \boldsymbol{t}^T ; \boldsymbol{u}^T]$. La technique proposée par Cootes pour trouver le meilleur jeu de paramètres consiste à déterminer les modifications optimales à apporter à $\boldsymbol{p}$ en fonction de la différence observée entre les niveaux de gris synthétisés $\boldsymbol{g_m}$ et les niveaux de gris de l'image réelle $\boldsymbol{g_r}$. Si l'on note $\boldsymbol{r}(\boldsymbol{p})$ ce résidu, nous avons donc :

$$\boldsymbol{r}(\boldsymbol{p}) = \partial \boldsymbol{g} = \boldsymbol{g_r} - \boldsymbol{g_m} \qquad \text{(eq. 1.9)}$$

On va chercher à minimiser l'erreur quadratique E :

$$\mathbf{E}(\boldsymbol{p}) = \boldsymbol{r}^T \boldsymbol{r} \qquad \text{(eq. 1.10)}$$

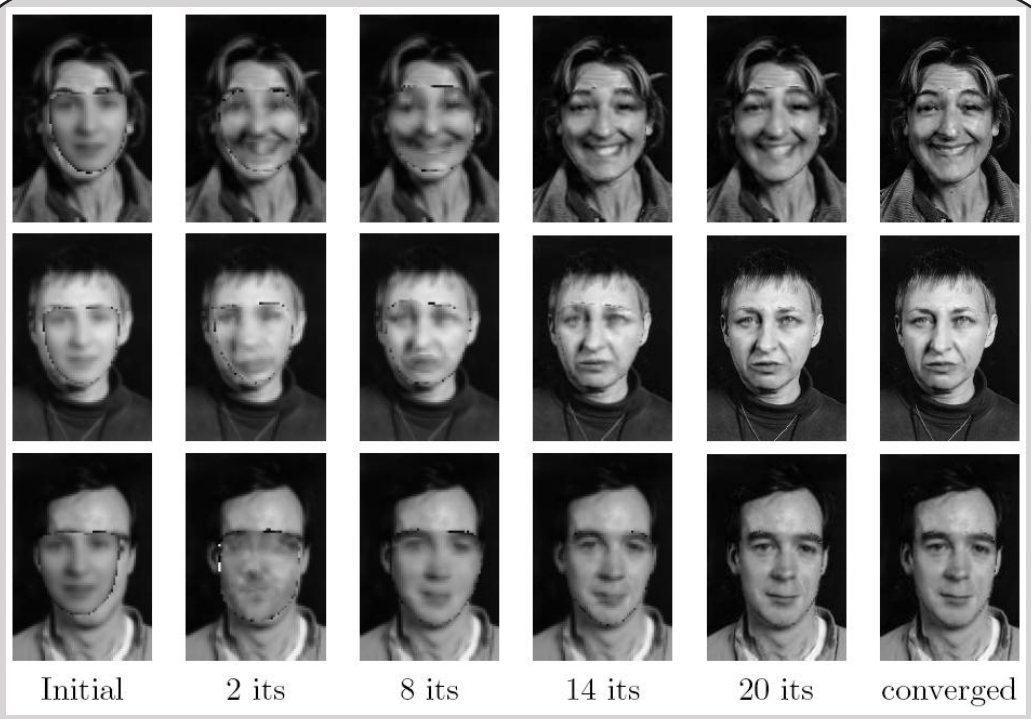


Figure 1.16 :
Exemple de convergence d'un AAM dans le cas d'un visage selon [Cootes, 1998].

Le développement de Taylor de l'équation 1.9 conduit à :

$$\partial \boldsymbol{p} = -\mathbf{R}\boldsymbol{r}(\boldsymbol{p}) \text{ avec } \mathbf{R} = \left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{p}}^T \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{p}} \right)^{-1} \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{p}}^T \quad \text{(eq. 1.11)}$$

Cootes a proposé de pré-calculer la matrice **R**, qui sera donc considérée comme constante. Cela peut être accompli en perturbant les vecteurs de paramètres optimaux connus de la base d'apprentissage et en enregistrant les d$\boldsymbol{p}$ et d$\boldsymbol{g}$ correspondants. La matrice **R** est alors obtenue en procédant à une régression linéaire à plusieurs variables. Une autre méthode pour l'obtenir est de remarquer que **R** peut être vue comme l'inverse de la matrice de Jacobi d'un algorithme de descente du gradient et peut donc alors être évaluée comme étant la moyenne de matrices de Jacobi calculées pour les images de la base d'apprentissage.

Cette approche, qui prend donc pour postulat que la relation entre les erreurs de reconstruction et les variations de paramètres est fixe et linéaire, donne des résultats satisfaisants pour des applications où l'éclairage est contrôlé et où les variations de poses sont limitées, pour un coût algorithmique limité. La figure 1.16 montre des exemples de convergence de l'AAM pour différents visages.

Raffinements des AAMs

Très peu de temps après l'article fondateur des AAMs, Cootes et al. ont proposé une première variation de l'algorithme original le Shape-AAM ([Cootes, 1998+]). Celui-ci est fort semblable dans le principe à l'AAM classique et ne diffère que sur la stratégie de convergence adoptée. Le résidu $\boldsymbol{r}$ étant toujours défini de la même façon, seul le modèle de forme et les paramètres de pose et d'échelle sont optimisés par une stratégie similaire à l'équation 1.11 :

$$\partial \boldsymbol{t} = \mathbf{R}_t \boldsymbol{r} \quad \text{et} \quad \partial \boldsymbol{b}_s = \mathbf{R}_s \boldsymbol{r} \qquad \text{(eq. 1.12)}$$

$\mathbf{R_S}$ et $\mathbf{R_t}$ étant de nouveau supposées fixes. Les paramètres du modèle de niveaux de gris ne sont donc pas optimisés mais sont directement estimés comme étant les valeurs modélisant le mieux l'image courante.

[Cootes, 2000] introduit une méthode pour appliquer la méthode de base de l'AAM au cas où le visage peut être vu sous différents angles : le multi-view AAM. En effet, l'hypothèse voulant que l'on considère **R** comme constante ne peut s'appliquer à des variations non-linéaires de l'apparence dues à des rotations du visage. La solution proposée est de construire cinq AAMs qui seront appliqués pour différentes poses possibles du visage (vue de face, vues de profil droit et gauche et deux vues intermédiaires).

Ensuite, dans [Cootes, 2001], les possibilités ont été élargies en introduisant les AAMs contraints. Cette variation des AAMs permet de prendre en compte des contraintes extérieures aux modèles qui sont soit fournies par d'autres algorithmes de segmentation opérant en parallèle, soit manuellement par un opérateur. En utilisant la théorie du Maximum de Vraisemblance, la convergence de l'AAM peut alors être guidée par des connaissances a priori sur la position de certains points ou sur la valeur de certains paramètres.

[Baker, 2001] a vu de son côté l'introduction de l'algorithme Inverse Compositional Image Alignment (ICIA). Le principal changement entre l'ICIA et l'optimisation proposée par Cootes est que la matrice de Jacobi n'est plus supposée être fixe et n'est donc pas apprise de manière définitive à partir de la base d'apprentissage. Elle est ici estimée en différenciant la fonction de coût par rapport aux paramètres du modèle de forme (qui seront les seuls à être optimisés comme pour le Shape-AAM) ce qui conduit à une matrice de Jacobi qui sera constante tout au long de la convergence mais spécifique à la série d'images traitée. Néanmoins, dans [Cootes, 2002] une comparaison est effectuée entre les performances du Shape-AAM et de l'ICIA, et si Cootes reconnaît que l'ICIA est élégante mathématiquement, il conclut que cela ne se traduit pas en un réel apport sur la performance. De même dans [Romdhani, 2005], l'ICIA est présentée comme une méthode intéressante mais dont les performances décroissent dès qu'elle est appliquée à des individus n'appartenant pas à la base d'apprentissage.

Dans un cadre 3D ([Dornaika, 2004]), il a été proposé d'utiliser une approche stochastique pour le tracking d'un AAM en faisant appel à la technique du filtrage particulaire ([Arulampalam, 2002]). L'optimisation des paramètres du modèle et la minimisation de l'équation 1.10 se fait alors en trois phases. La première est une descente

du gradient qui fournit une première solution intermédiaire. Le problème étant à grande dimension, l'algorithme de descente du gradient est cependant susceptible de converger sur un minimum local. La seconde phase, la partie stochastique du procédé, va consister à procéder à une diffusion de la solution intermédiaire afin d'obtenir une nouvelle solution intermédiaire correspondant au Maximum de Vraisemblance. La troisième phase consiste en une nouvelle descente du gradient qui fournira la solution finale.

Dans le même cadre des problèmes 3D où la linéarité de la matrice de Jacobi ne peut donner de résultats satisfaisants, [Romdhani, 2005] a, quant à lui, proposé une technique d'optimisation complexe à caractéristiques multiples (Multiple Features Fitting Strategy, MFFS). Partant de la constatation que la convergence vers le minium optimal d'une fonction de coût peut être un problème ardu à cause des minimums locaux, l'idée force de la MFFS est l'utilisation en parallèle de plusieurs fonctions de coût qui seront fusionnées afin d'obtenir une fonction de coût global qui sera plus robuste. Cette approche est similaire aux stratégies utilisées pour la classification d'objet comme le boosting ([Schapire, 1997]) où des classifieurs forts sont formés en combinant des classifieurs faibles. Trois caractéristiques d'images (les niveaux de gris, le gradient, la réflexion des

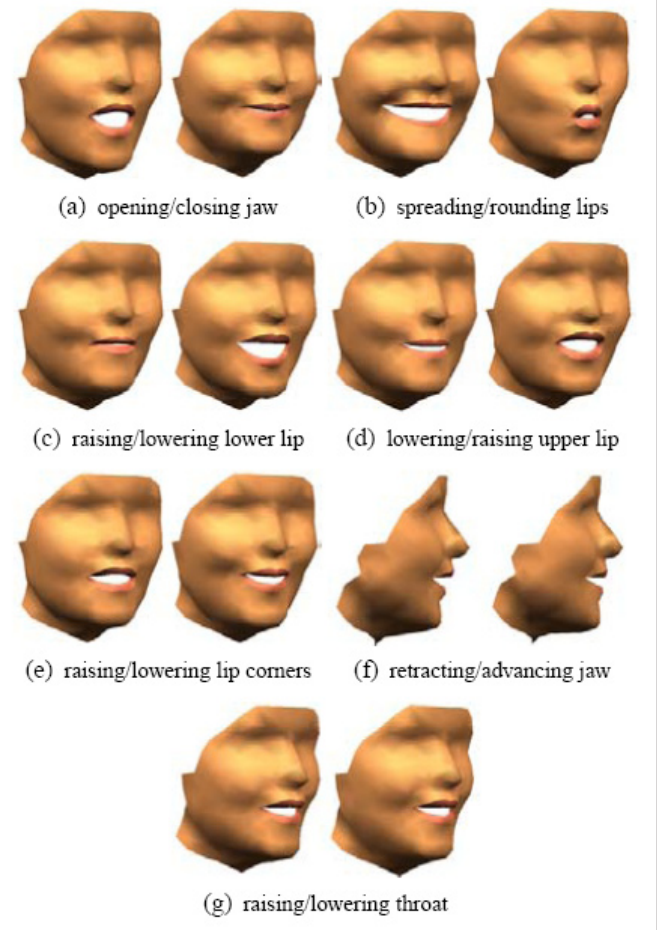


Figure 1.17 :
Modes de variation d'un modèle actif d'apparence bâti de sorte à ce que chaque mode variation est une interprétation articulatoire et donc phonétique, selon [Odisio, 2005].

sources lumineuses) et deux caractéristiques de modèles (la distribution des points d'un modèle 3D et des contraintes de textures) sont utilisées dans un procédé de fusion afin d'obtenir la segmentation finale.

Parmi les très nombreux travaux sur les AAMs appliqués au visage humain, nous pouvons encore citer parmi les approches originales [Daubias, 2002] où est proposée une construction semi-automatique du modèle d'apparence. Les mouvements labiaux sont d'abord enregistrés alors que le locuteur a été préparé : ses lèvres ont été teintes en bleu de façon à capturer parfaitement la forme, ce qui permet de construire le PDM. Un deuxième enregistrement est ensuite effectué, sans maquillage bleu, durant lequel le locuteur doit effectuer les mêmes mouvements que précédemment. Ainsi, la forme étant censée être connue, on peut alors prélever l'apparence correspondante. La principale limitation de cette stratégie est que les mouvements labiaux entre les deux enregistrements ne pourront jamais être parfaitement identiques ce qui entraînera des erreurs sur le modèle statistique, même si les résultats présentés suggèrent que cette imprécision est acceptable.

Dans [Odisio, 2003], les auteurs présentent une stratégie dans laquelle le modèle statistique est construit de façon guidée dans une optique articulatoire. Plusieurs ACPs sont effectuées consécutivement à chaque fois sur un nombre restreint de points du PDM du visage, les contributions des points à la variance totale étant retirées à chaque étape, ce qui conduit à un modèle de forme où les modes de variations correspondent tous à un mouvement articulatoire totalement déterminé, ce qui présente un intérêt évident pour synthétiser les mouvements de la parole (voir figure 1.17). Ces modèles de forme (et les modèles d'apparence qui en découlent) sont néanmoins exclusivement mono-locuteur et demandent en outre une longue étape de construction.

Enfin, dans [Gross, 2004], il est proposé une version des AAMs permettant de tenir compte des problèmes d'occultation d'une partie du visage en rajoutant des images présentant différents cas de figures d'occultations dans la base d'apprentissage.

1.3 Bilan

Les parties précédentes ont présenté de nombreuses méthodes et approches pour effectuer la segmentation des lèvres.

Les approches pixels donnent des résultats manquant de finesses pour les contours des lèvres mais peuvent néanmoins fournir des informations intéressantes sur la zone labiale. Dans le cadre de notre travail, une méthode appartenant à cette famille sera utilisée comme prétraitement afin d'obtenir une classification grossière des pixels en lèvres ou peau grâce à des critères couleurs.

Les approches par modèle analytique donnent des résultats satisfaisants pour de nombreuses images. Si leur flexibilité est gage d'une large applicabilité, la contrepartie est que les modèles peuvent converger sur des formes non-réalistes dans certains cas de figure. En outre, la phase de paramétrisation peut être très longue.

Si les modèles actifs ne nécessitent pas de paramétrisation empirique (le modèle étant construit directement par analyse statistique), une phase d'annotation des données est cependant nécessaire. Sous l'hypothèse que la base d'apprentissage est suffisamment

grande, les modèles actifs permettent de générer facilement des formes réalistes de lèvres.

Les parties suivantes de ce rapport vont présenter notre propre déclinaison des principes des ASMs et des AAMs dans l'optique de créer un modèle multi-locuteurs présentant un bon compromis robustesse/complexité. L'approche sera de développer la méthode dans le cadre plus simple d'un locuteur unique puis de rendre les principes de base transposables à une variabilité de la personne traitée.

Nous allons présenter deux façons de décrire l'apparence (qui dans cette thèse désignera les valeurs prises dans l'espace couleur YCbCr) : en ayant recours à une grille d'échantillonnage pour prélever les valeurs des pixels et en utilisant des descripteurs locaux du voisinage des points de contrôle de la forme.

Une large partie sera consacrée à la conception d'une fonction de coût adaptée à la zone de l'intérieur de la bouche et ses non-linéarités (ouverture/fermeture, présence/absence des dents...).

L'état de l'art a en outre présenté diverses méthodes d'optimisation des paramètres des modèles, dont certaines techniques de convergence des fonctions de coût relativement complexes adaptées par exemple au cas 3D. Outre le fait que notre travail s'est limité à un cadre 2D, les centres d'intérêt de notre thèse sont davantage la modélisation statistique des données que les méthodes algorithmiques d'optimisation, si bien que nous avons opté pour une stratégie classique de descente du Simplex en procédant néanmoins à des étapes d'initialisation afin d'améliorer vitesse et robustesse.

Enfin, la qualité d'une segmentation est généralement évaluée par des critères quantitatifs mesurant l'écart entre les points de contrôle détectés et leurs positions réelles (connues car annotées manuellement). Une telle évaluation est objective mais ne dit néanmoins pas si la segmentation a été capable de restaurer les mouvements labiaux avec suffisamment de réalisme pour tirer partie de la nature multimodale de la parole. Nous procéderons donc dans ce rapport à une évaluation subjective en testant l'intelligibilité d'un clone synthétisé à partir de l'analyse dans un cas mono-locuteur. Pour cela, l'apport pour la compréhension de numéro de téléphone de notre schéma d'analyse/synthèse a été mesuré.

CHAPITRE 2

Modèle Mono-Locuteur

Ce chapitre traite de notre méthode pour segmenter les lèvres dans un cas mono-locuteur. Nous présentons d'abord le cadre du problème et les espaces couleurs et prétraitements utilisés pour représenter les images à traiter. Par la suite, un modèle local de commissures des lèvres, qui servira lors de l'initialisation de la segmentation, est introduit ainsi que des Modèles Actifs de Forme et d'Apparence de la zone labiale. Enfin, nous présentons et comparons diverses fonctions de coût permettant d'optimiser les paramètres des modèles ainsi qu'un schéma d'analyse adapté au problème.

2.1 Cadre du problème

2.1.1 Applicabilité du modèle

Dans notre travail, nous développerons des modèles bidimensionnels aptes à traiter des images représentant des visages vus de face et ne tolérant donc que de légères rotations du visage du sujet.

Nous considérons en outre soit que le visage a été localisé lors d'un pré-traitement, soit que le cadre de l'image est fixe et connu, et donc qu'en tout état de cause l'on connaît assez grossièrement la position de la bouche et de son voisinage (c'est à dire que l'image est cadrée sur la partie inférieure du visage).

Figure 2.1 :
Locuteur filmé par micro-caméra.

Une abondante littérature traite du problème de la détection du visage humain. Entre autres exemples, dans [Rowley, 1996], un réseau de neurones est utilisé pour classer de petites fenêtres d'images comme étant des visages ou non. Dans [Yang, 1998], les auteurs modélisent la distribution colorimétrique du visage et une segmentation multi-échelle, La Cascia et al proposent d'avoir recours à une carte de texture ([La Cascia, 1998]), [Sugimoto, 2004] introduit des critères de forme et de couleurs pour détecter l'ellipse du visage grâce à quelques traits saillants. Enfin, dans [De Otalera, 2005], Modèles de Markov Cachés et contours actifs sont combinés pour accomplir la détection.

Dans le cas de cette étude mono-locuteur, nous sommes dans la configuration où les images de la base de données ont été filmées par une micro-caméra solidaire du visage du locuteur (comme sur la figure 2.1).

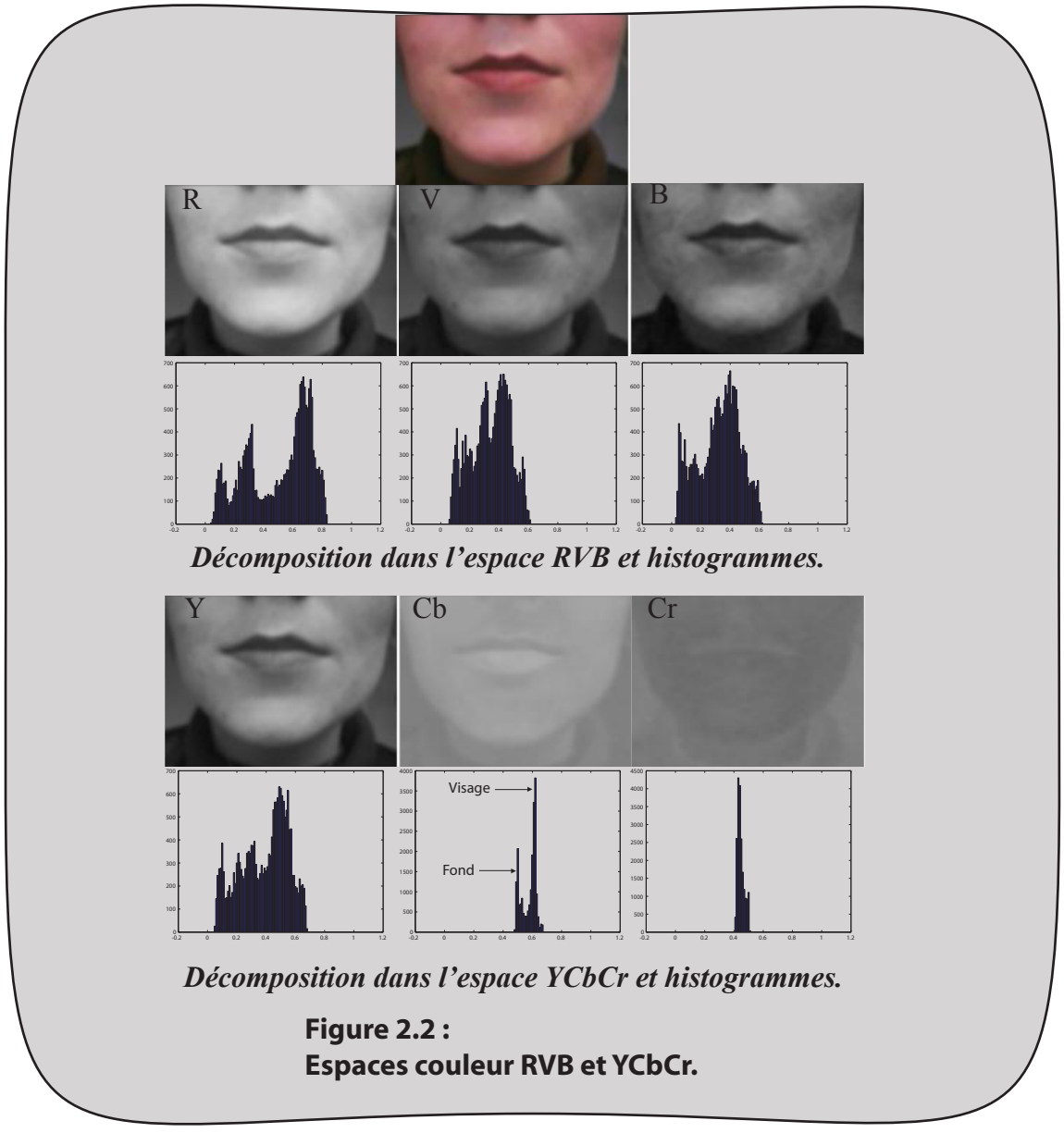


Figure 2.2 :
Espaces couleur RVB et YCbCr.

2.1.2 Espaces couleur

Comme les informations de colorimétrie et de luminance sont mélangées dans l'espace couleur RVB, nous avons choisi d'utiliser l'espace YCbCr où les composantes chromatiques CbCr sont séparées de la luminance.

Le passage entre les deux espaces couleur est bijectif et s'effectue grâce aux formules suivantes :

$$\begin{aligned} Y &= 0.299*R+0.587*V+0.114*B \\ Cb &= \left(-0.169*R-0.331*V+0.500*B\right)+0.5 \\ Cr &= \left(0.500*R-0.419*V-0.081*B\right)+0.5 \end{aligned} \qquad \text{(eq. 2.1)}$$

les termes +0.5 servent simplement à recadrer les valeurs de Cb et Cr entre 0 et 1 (au lieu de -0.5 et +0.5).

La figure 2.2 montre la décomposition dans les espace RVB et YCbCr d'une même image ainsi que les histogrammes correspondants. On constate que les composantes chromatiques Cb et Cr ont des distributions de valeurs relativement concentrées. La composante Cb présente ici un pic très distinct correspondant à la colorimétrie du visage tandis que la séparation entre le visage et le fond est plus ardue sur Cr dans le cas présenté.

Nous utiliserons également dans ce rapport la pseudo-teinte **H** proposée dans [Poggio, 1998]. L'avantage de cette caractéristique est qu'elle présente un fort contraste entre les lèvres et la peau, ce qui correspond à un vecteur gradient de grande norme au niveau du contour. **H** est calculée en un point (x,y) comme :

$$H(x,y)=\frac{R(x,y)}{R(x,y)+V(x,y)} \qquad \text{(eq. 2.2)}$$

où R et V sont les composantes RVB mais **H** peut également aisément être exprimée en fonction des composantes YCbCr.

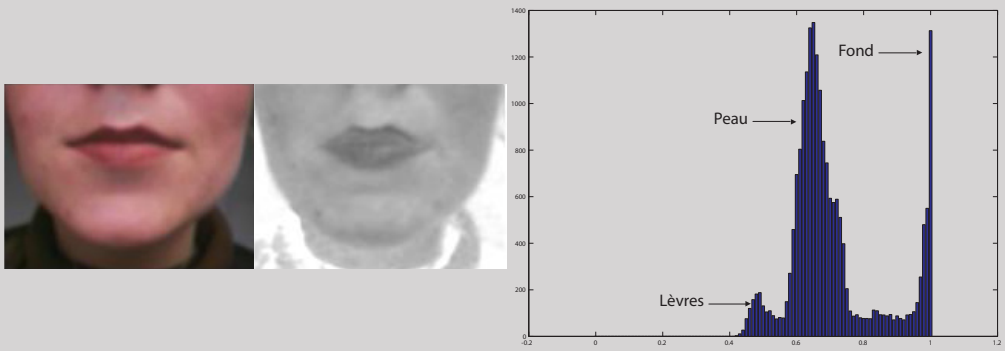


Figure 2.3 :
Pseudo-teinte, exemple et histogramme.

La figure 2.3 présente la pseudo-teinte d'une image ainsi que son histogramme sur lequel il est facile d'identifier la peau, les lèvres et le fond de l'image. Il est à noter que la séparation observée ici entre la pseudo-teinte des lèvres et celle de la peau est robuste lorsque l'on change de locuteur.

2.1.3 Un pré-traitement intéressant : le filtrage rétine

L'un des écueils les plus importants de toutes les applications de traitement d'images est la sensibilité au changement d'éclairage. Dans le cadre de l'analyse labiale, un changement d'illumination fait que les valeurs des pixels peuvent changer même si le locuteur est immobile.

Si l'on veut supprimer ou au moins réduire cette variabilité, en supposant que les composantes CbCr sont suffisamment insensibles à ces changements, il convient alors de faire un traitement adapté pour diminuer la sensibilité de la composante Y. Ceci peut être accompli en effectuant un pré-filtrage inspiré du modèle biologique de la rétine humaine ([Beaudot, 1994]). Le filtre rétinien possède de nombreuses propriétés intéressantes pour le prétraitement spatio-temporel d'une image qui ont par exemple été exploitées récemment pour l'analyse de mouvement ([Benoit, 2005]) ou la segmentation d'indices faciaux ([Hammal, 2004]).

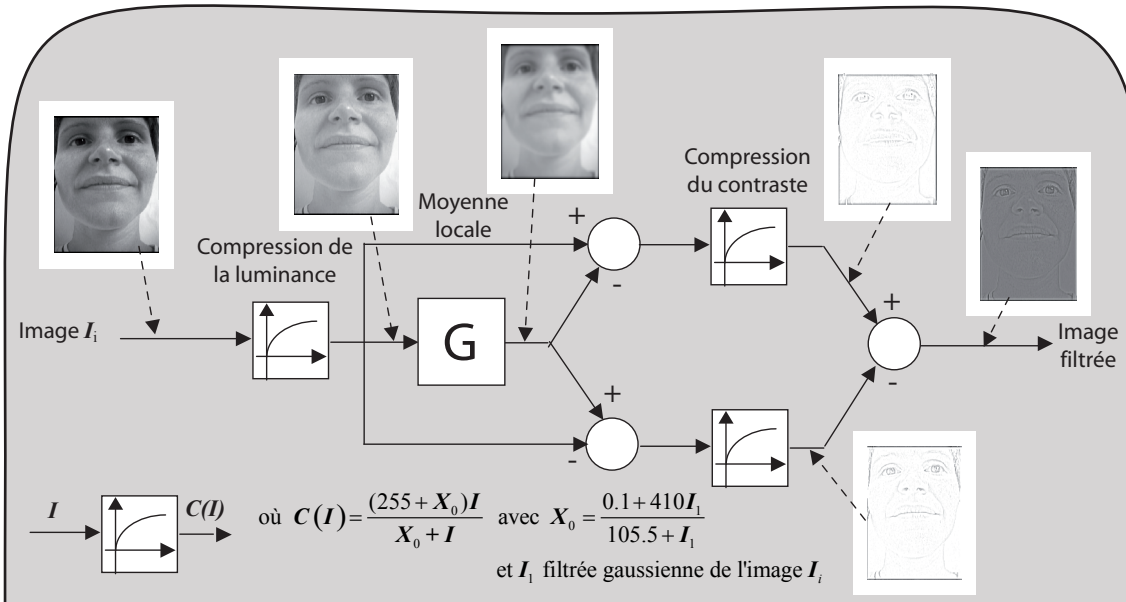


Figure 2.4 :
Schéma de principe du filtrage rétine.
L'image de luminance I_i subit d'abord une compression locale puis passe à travers un filtre gaussien G. On effectue ensuite deux différences opposées entre la luminance compressée et sa filtrée. Les deux résultats intermédiaires sont ensuite de nouveau compressés puis on obtient l'image résultat en calculant leur différence.

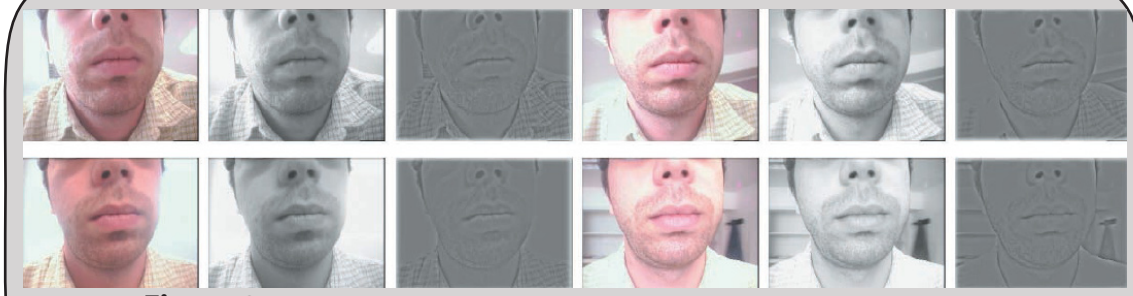


Figure 2.5 :
Illustration du filtre rétine (image, luminance et luminance filtrée) pour la même personne sous quatre éclairages différents.

Dans ce travail nous n'utiliserons que les capacités de filtrage spatial, ce qui correspond à la modélisation de fonctionnement de la couche plexiforme externe de l'oeil. Ainsi, son effet sera de rehausser les contours (effet analogue à un filtre passe-bande) et d'atténuer les variations locales d'illuminations. Le schéma de principe du filtrage rétinien est donné par la figure 2.4.

La figure 2.5 illustre son effet sur une séquence d'images avec d'importantes variations d'éclairage. Si la luminance subit d'importantes fluctuations, sa filtrée reste en grande partie constante sur la séquence.

2.2 Base d'images d'apprentissage

Nous disposons de 40 séquences vidéos (en 25Hz détramées) de la même personne prononçant des numéros de téléphone à dix chiffres (voir figure 2.6). Il est à noter que chaque numéro de 0 à 99 est présent au moins une fois dans les séquences (voir Annexe 1). Dans la moitié de ces vidéos le locuteur murmure tandis que dans l'autre moitié son élocution est normale. Ces vidéos durent environ 8 secondes chacune, ce qui représente au total quelques 8000 images dont un certain nombre seront manuellement annotées.

Annoter manuellement une image consiste à saisir des points de contrôle qui définiront les contours internes et externes des lèvres ainsi que ceux des dents.

Nous avons décidé de décrire la forme générale de la zone de la bouche par 30 points de contrôle : 12 sur le contour extérieur des lèvres et 8 sur le contour intérieur. Enfin les contours des dents sont décrits par 10 points supplémentaires (voir la figure 2.7).

Le choix des points de contrôle à annoter est intuitif : pour le contour externe supérieur des lèvres, nous avons les commissures, trois points sur l'arc de Cupidon et deux points qui permettront de caractériser la courbure du contour. Pour le contour externe inférieur nous avons, outre les commissures, un point caractéristique de l'ouverture et quatre caractérisant la courbure (contre deux sur le contour supérieur, du fait d'une longueur curviligne plus importante des courbures du contour inférieur).

Pour le contour intérieur nous avons des points sur les "commissures internes", des points pour caractériser l'ouverture et enfin quatre points pour modéliser plus fidèlement la courbure du contour.

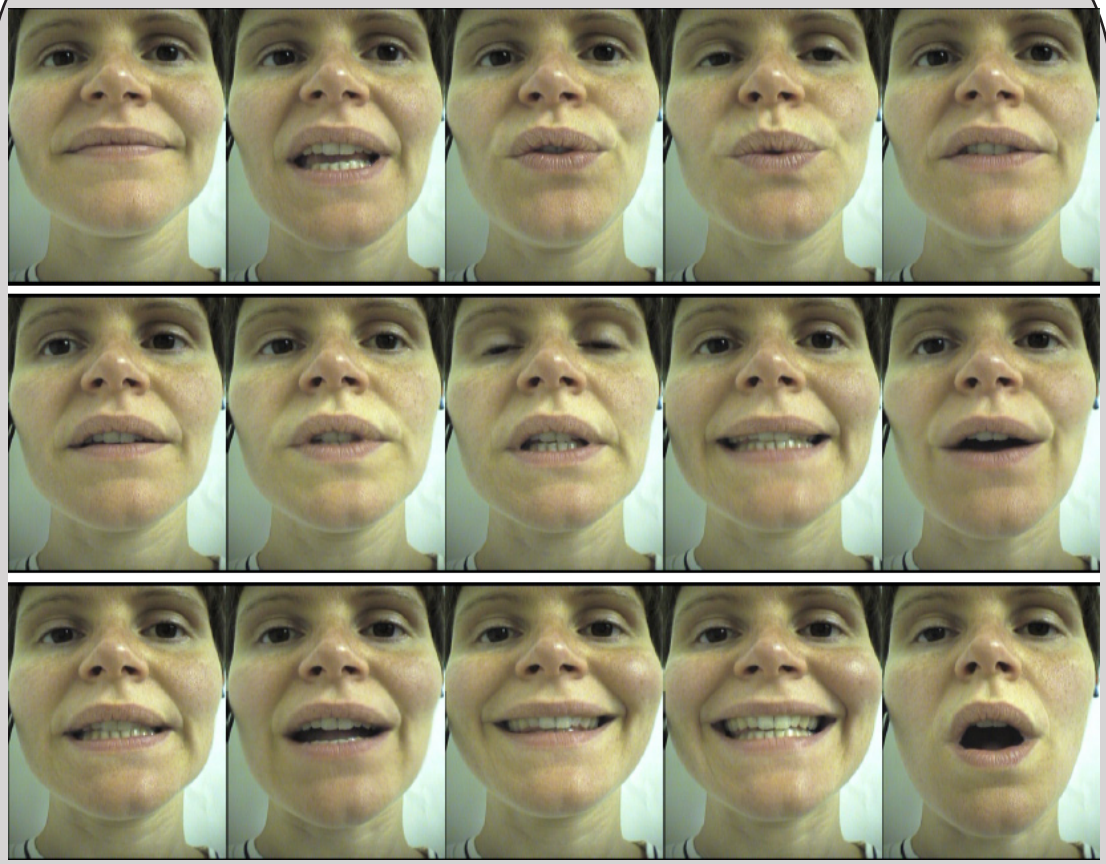
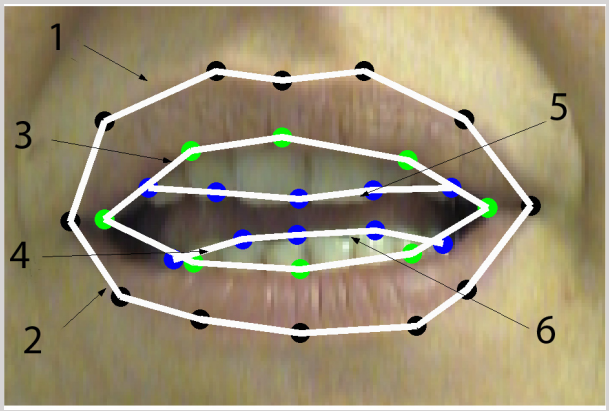


Figure 2.6 :
Exemples d'images de la base de données.



Courbes 1 et 2: contour extérieur de la bouche, 3 et 4: contour intérieur de la bouche, 5 et 6: contours des dents.

Figure 2.7 : Exemple d'annotation manuelle d'une image de la base d'apprentissage.

Figure 2.8 :
Image de référence de bouche au repos ainsi que sa symétrique verticale.

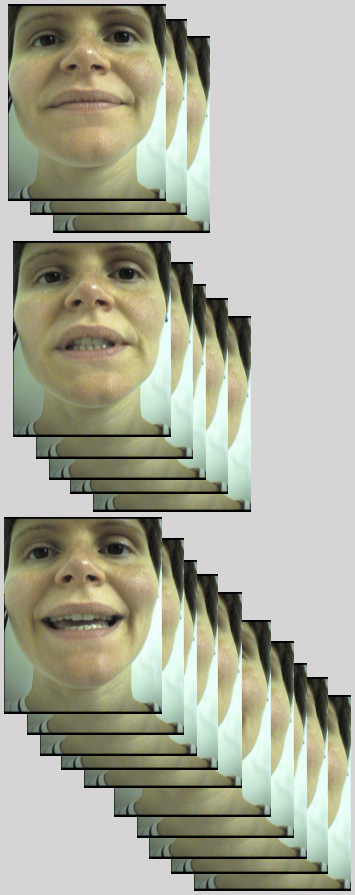
De la même façon les dents sont définies par dix points, cinq pour chaque rangée. Dans le cas où la bouche est fermée et plus généralement lorsque les dents n'apparaissent pas, les points correspondants sont confondus avec ceux du contour intérieur des lèvres.

200 images de la base de donnée ont été sélectionnées pour faire partie de la base d'apprentissage sur laquelle les différents modèles seront entraînés. Les images en question ont été choisies de façon à offrir la plus large variété possible de formes de bouche. En outre, leurs symétriques verticales ont également été inclues, afin d'avoir un modèle pouvant gérer diverses direction d'éclairage (voir figure 2.8), si bien que N=400 images constitueront la base de données. Enfin, une image où la bouche est fermée et "au repos" a été sélectionnée comme référence (voir figure 2.8).

Les coordonnées des points de contrôle sont sauvegardées dans les vecteurs $\boldsymbol{X}_i$ ($1 \leq i \leq N$) qui constitueront notre vérité terrain.

Enfin, cinq vidéos complètes (soit plus de 900 images, 1800 avec leurs symétriques) ont également été sélectionnées et étiquetées afin de servir à tester la robustesse de l'algorithme de façon quantitative. Les images restantes servant à des évaluations qualitatives de la méthode (voir figure 2.9).

Les coordonnées des points de contrôle manuellement étiquetés doivent ensuite être normalisées : la commissure gauche est ramenée au point (0,0) puis les coordonnées sont divisées par la largeur de la bouche de référence (c'est à dire la distance inter-commissures, ce qui est similaire aux FAP du MPEG4-SNHC, voir 1.1.2). Contrairement à la transformation procrustéenne généralisée ([Cootes, 1998]), le centrage ne s'effectue donc pas sur le centre de gravité des points de contrôle, ceci afin de faciliter l'utilisation des commissures comme points-clés pour placer les lèvres sur les images à traiter. Il serait en outre possible de normaliser pour chaque image les coordonnées des points par la largeur de la bouche courante mais cela conduit à décorréler certains mouvements labiaux de leur effet sur la largeur de la bouche (on notera que la phase d'homothétie de la transformation procrustéenne d'homothétie conduirait aussi à une telle décorrélation). Par exemple, une bouche "en O" se traduit par une diminution de la distance entre les deux commissures alors qu'un large sourire conduit au contraire à leur écartement. Enfin, on peut remarquer que cette normalisation de la largeur pourrait ne pas être effectuée dans ce chapitre, la cadrage du visage étant constant dans toutes les séquences de notre locuteur. Néanmoins, cette normalisation sera indispensable au chapitre 3 où l'on traitera du cas multi-locuteurs avec des conditions d'acquisition variables et nous la faisons ici par homogénéité.



Base d'apprentissage de 200 images, 400 avec les symétriques.

Base de test d'environ 1800 images (cinq séquences complètes plus leurs symétriques) manuellement annotées qui serviront principalement aux évaluations quantitatives (et à l'évaluation qualitative au chapitre 4).

Reste des images de la base de données (environ 7000 images) qui serviront à des évaluations subjectives au chapitre 4.

Figure 2.9 :
Répartition des images de la base de données.

Finalement, les 30 points de contrôle normalisés correspondent donc à 2x29=58 degrés de libertés.

Lorsque l'on voudra segmenter les lèvres sur une image inconnue, il sera nécessaire de déterminer 3 degrés de liberté supplémentaires : la position $\boldsymbol{t}$ de la commissure gauche (ses coordonnées) ainsi que l'échelle q qui permettra d'adapter le modèle à la taille de l'image (q correspondant donc à la largeur de la bouche de repos à l'échelle de l'image traitée).

Enfin, simultanément à l'étiquetage de la base d'apprentissage, l'opérateur affecte également à chaque image un "Etat Général de la Bouche" (EGB). L'EGB est une variable d'état qui décrit de manière élémentaire la configuration de la bouche en la classant dans une classe typique. L'EGB a vocation à améliorer la rapidité, la précision et la robustesse de la convergence en sélectionnant la meilleure initialisation possible pour l'optimisation du modèle.

Après avoir déterminé empiriquement les types d'images les plus susceptibles d'être mal segmentées, le nombre d'EGB a été fixé en conséquence à 4 : 1) bouche fermée, 2) bouche ouverte, 3) bouche souriante, 4) bouche grande ouverte (par exemple à cause de la peur ou de la surprise). Ces quatre EGB permettent d'obtenir un bon pavage de l'espace des réalisations et un exemple de chaque EGB est présenté sur la figure 2.10.

Figure 2.10 :
Exemples d'images typiques pour chaque EGB.
De gauche à droite: bouche fermée, ouverte, souriante, bouche en O.

$$\text{pour } 1 \le i \le N,\ EGB(i) = \begin{cases} 1 \text{ si la bouche est fermée} \\ 2 \text{ si la bouche est ouverte} \\ 3 \text{ si la bouche est souriante} \\ 4 \text{ si la bouche est en O} \end{cases} \quad \text{(eq. 2.3)}$$

Les catégories 3 et 4 correspondent à des images qui sont en général plus susceptibles de mettre à mal la précision de l'algorithme. En effet, la configuration de la bouche est alors relativement éloignée de la forme moyenne ce qui correspond à des valeurs de paramètres plus extrêmes.

Il est à noter que certaines images sont difficiles à classifier dans un EGB ou un autre et ce même pour un opérateur humain, ce qui laisse donc à penser que cela sera difficile à faire automatiquement de façon robuste. Si nous montrons par la suite que déterminer l'EGB améliore la vitesse et la précision de la convergence, une mauvaise classification n'est cependant pas nécessairement rédhibitoire pour la qualité finale de la segmentation. Enfin, les images se trouvant à la «limite» entre deux EGB correspondent donc à des configurations de bouche non extrêmes et donc à des valeurs de paramètres a priori facilement atteignables.

2.3 Modèles

2.3.1 Modélisation locale des commissures des lèvres

2.3.1.1 Principe

Les commissures des lèvres sont des points-clés dont la connaissance permet de déterminer la position et l'échelle de la bouche.

Ces points sont particulièrement difficiles à détecter du fait qu'ils se trouvent sur un contour généralement flou et dans une zone ombragée (comme le montre la figure 2.11)

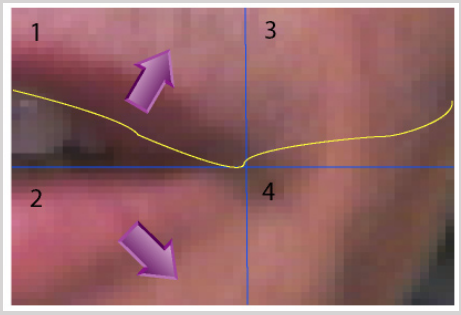


Figure 2.11 :
Zone des commissures des lèvres avec les régions caractéristiques, la direction des vecteurs gradients et la ligne des minima de luminance.
Les zones 1 et 2 correspondent aux frontières entre les lèvres et la peau et donc à des gradients marqués. Les zones 3 et 4 sont à l'opposé homogènes.

et les critères classiques de détection de points d'intérêt s'appliquent donc mal à ce cas. Même le placement manuel de ces points peut se révéler difficile si le voisinage des points de commissures est réduit : un opérateur a en effet tendance à prolonger les contours des lèvres pour placer les points. Cet état de fait a par exemple été exploité dans [Reveret, 1999] où les commissures sont déterminées comme étant l'intersection des contours inférieurs et supérieurs des lèvres.

Nous faisons la même hypothèse que dans [Delmas, 2002] et [Eveno, 2004] en supposant que les commissures des lèvres se trouvent sur la ligne reliant les minima de luminance pour chaque colonne de l'image. Cela supprime un degré de liberté et il suffit donc de trouver l'indice des colonnes pour chaque commissure.

Cette ligne des minima est tracée en deux étapes.

Tout d'abord on détermine un point initial, ou germe. Pour cela on calcule pour chaque colonne la ligne du minimum. La figure 2.12 montre que certains points peuvent se retrouver dans les cavités nasales ou sur des lignes arbitraires dès lors que la colonne ne comprend plus la zone de la bouche. Nous calculons alors un accumulateur pondéré des indices de lignes pris par tous ces points en favorisant les points proches du centre de l'image et en pénalisant les autres (entendu que le visage a été détecté en pré-traitement et

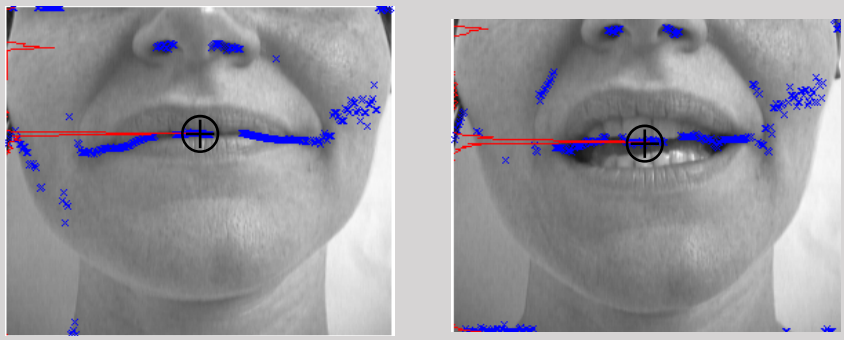


Figure 2.12 :
Exemple de détermination du germe (pour la construction de la ligne des minima de luminance.
+ noir : germe, croix bleues : minima pour chaque colonne, tracé rouge : valeur de l'accumulateur pour chaque ligne.

que l'image est donc plus ou moins centrée sur la bouche). La procédure en question est la suivante :

$$\begin{aligned}
&pour\ ind_col = 1{:}largeur \\
&\quad ind_lin = \text{indice du mininum}\left(\text{Image}\left(ind_col\right)\right) \\
&\quad \text{accumul}\left(ind_lin\right) = \text{accumul}\left(ind_lin\right) + \frac{1}{\sqrt{\left(ind_col - \frac{largeur}{2}\right)^2 + \left(ind_lin - \frac{hauteur}{2}\right)^2}} \\
&fin
\end{aligned} \qquad \text{(eq. 2.4)}$$

où *largeur* et *hauteur* désignent la taille de l'image traitée et *ind_lin* et *ind_col* les indices des lignes et colonnes. La ligne récoltant le meilleur score est donc sélectionnée pour le germe, la colonne retenue étant la centrale. La figure 2.12 montre le tracé de la valeur de l'accumulateur, la pointe du pic désignant le germe.

Une fois le germe connu, la ligne est tracée dans chaque direction en se propageant de colonne en colonne grâce à un algorithme de suivi du minimum de luminance.

La ligne des points d'intérêt étant connue, nous avons adopté comme approche le fait de considérer que les commissures sont à l'intersection de quatre régions de caractéristiques différentes. En se référant à la figure 2.11, les régions 1 et 2 sont non-homogènes car caractérisées par une frontière entre les lèvres et la peau. Les régions 3 et 4 sont, quant à elles, homogènes d'un point de vue chromatique mais peuvent présenter des différences de luminance.

Chacune des régions va alors être décrite par un jeu de descripteurs locaux. Nous avons décidé d'utiliser les filtres de dérivées de gaussiennes qui, après avoir été introduits dans [Marr, 1980], ont été utilisés avec succès dans divers problèmes de traitement d'image ([Young, 1985], [Young, 1995], [Lindeberg, 1998], [Hall, 2000]) dont la détection de points d'intérêt ([Schmid, 2000]) et la détection de visage ([Vogelhuber, 2000]), voir figure 2.13.

Dans un cas unidimensionnel, les filtres sont calculés de la façon suivante :

$$\begin{aligned}
g_0(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\
g_n(x) &= \frac{d^n g_0(x)}{dx^n} = \frac{1}{\sqrt{2\pi\sigma}} \left(\frac{1}{\sigma}\right)^n He_n\left(\frac{x}{\sigma}\right)
\end{aligned} \qquad \text{(eq. 2.5)}$$

où He_n est le polynôme d'Hermite de rang n.

R.A. Young a par exemple proposé de les utiliser pour modéliser les champs réceptifs du système visuel humain ([Young, 1985]). L'intérêt de ces filtres est leur capacité à coder de façon pertinente l'information visuelle utile à la perception humaine.

Ils procurent une base orthonormale permettant de décrire un signal de façon optimale. Plus le nombre de dérivées utilisées est important et plus la description du signal est précise.

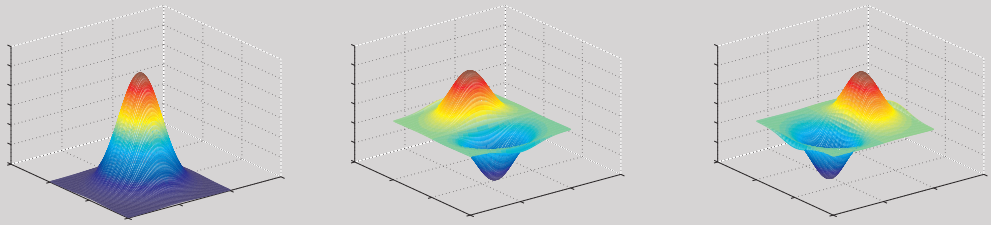


Figure 2.13 :
Filtres dérivés de gaussiennes G, Gx, Gy : moyenne et premières dérivées dans chaque direction.

Dans notre cas, les filtres seront évidemment bidimensionnels et nous nous limiterons à la première dérivée dans chaque direction (figure 2.13).

En pratique, ces filtres sont des fenêtres de convolutions dont la taille est déterminée par la taille de la zone du visage (figure 2.14). Nous avons déterminé empiriquement qu'une taille adaptée à l'utilisation souhaitée correspondait à environ un dixième de la largeur de la bouche. En considérant le visage détecté dans un prétraitement, une taille de filtre pertinente est alors égale à un vingtième de la largeur du visage.

Comme nous limitons le modèle aux premières dérivées, nous devrons finalement calculer les convolutions entre trois fenêtres correspondant aux filtres G, Gx et Gy (respectivement la moyenne, le gradient horizontal et le gradient vertical) pour les trois composantes YCbCr.

Enfin, les flltres sont centrés sur les zones qu'ils décrivent : par rapport au point de commissure que l'on veut modéliser, les filtres sont décalés de la moitié de leur taille vericalement et horizontalement (décalage positif ou négatif selon la zone décrite).

La position des commissures étant connue sur la base d'images étiquetées, on peut procéder à un apprentissage : pour chaque image les réponses des filtres sont stockées dans des vecteurs à 36 composantes (3 filtres x 3 composantes x 4 régions).

Ces données vont nous permettre d'évaluer la distribution statistique de ces vecteurs et de bâtir deux modèles aux fonctions complémentaires.

Le premier modèle testera si un pixel est susceptible d'être un point de commissure gauche ou droite (ou aucun des deux s'il est en dessous d'un seuil de probabilité). Le second testera ensuite si un couple de pixels est un couple de commissures (ainsi il est construit pour traiter des vecteurs de 36x2=72 composantes). En testant les couples formés par les meilleurs pixels candidats désignés par la première étape, il augmente donc la robustesse de la méthode en garantissant une cohérence entre les deux commissures détectées.

Les vecteurs de données à traiter étant de grandes tailles, on procède à des ACPs pour diminuer l'espace considéré. En gardant 95% de la variance, on obtient pour chacun des deux modèles des données réduites à 6 valeurs.

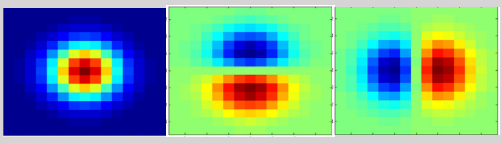


Figure 2.14 :
Fenêtres de convolutions correspondant aux filtres dérivés gaussiens G, Gx et Gy

2.3.1.2 Modèle par mélange de gaussiennes et algorithme EM

La modélisation statistique retenue pour les deux modèles de commissures est un mélange de gaussiennes.

Le premier modèle permettra de classer un pixel dans l'une des deux catégories y_i possibles ($i = \{1,2\}$), avec y_1: commissure gauche et y_2: commissure droite. Le second modèle n'a en revanche qu'une seule classe : le couple de pixels testé est un couple de commissures ou non.

Ces modélisations sont adaptées à notre tâche puisque celles-ci créent des sous-classes susceptibles de rendre compte de la variabilité de l'apparence des commissures selon la configuration de la bouche (ouverte, fermée, etc...). Nous allons détailler dans la suite la méthode pour construire le premier modèle, le second étant obtenu de façon analogue (à la différence, donc, qu'il n'a qu'une seule classe).

Les mélanges de gaussiennes G_i correspondant à chacune des deux classes y_i sont définies par l'ensemble des K_i gaussiennes $\{G_{i,j}\}_{j=1...K_i}$. Chacune de ces gaussiennes est définie par les paramètres ($w_{i,j}$, $\boldsymbol{\mu}_{i,j}$ et $\boldsymbol{\Sigma}_{i,j}$) qui correspondent respectivement au coefficient de poids associée à la gaussienne, à sa moyenne et à sa matrice de covariance.

Le nombre de gaussiennes K_i est déterminé grâce au principe de la Minimum Description Length ([Carson, 2002]). Pour un nombre K_i de gaussiennes, leurs paramètres optimaux (variance, moyenne, amplitude) sont déterminés itérativement grâce à l'algorithme EM (estimation/maximisation, [Carson, 2002]).

Finalement le nombre de gaussiennes retenu est celui qui correspond au meilleur équilibre entre la simplicité et la précision de la modélisation des données d'apprentissage (principe du rasoir d'Ockham qui veut que le modèle le plus simple pour représenter un système est le meilleur). Les modèles les plus complexes (c'est à dire ceux correspondant à un nombre élevé de gaussiennes) sont donc pénalisés.

Soit $\boldsymbol{W}$ un vecteur contenant les réponses des descripteurs locaux (réponses dont la taille a été diminuée par ACP pour revenir à 6 valeurs). Pour un nombre donné de gaussiennes K_i, la densité de probabilité d'appartenance d'un vecteur $\boldsymbol{W}$ à l'une des deux catégories est alors :

$$p(\boldsymbol{W} \mid y_i) = \sum_{j=1}^{K_i} w_{i,j} \times p\left(\boldsymbol{W} \mid \left(\boldsymbol{\mu}_{i,j}, \Sigma_{i,j}\right)\right) \qquad \text{(eq. 2.6)}$$

où K_i est le nombre de gaussiennes modélisant la catégorie et où $w_{i,j}$, $\boldsymbol{\mu}_{i,j}$ et $\boldsymbol{\Sigma}_{i,j}$ les caractéristiques des gaussiennes (respectivement poids, moyenne et matrice de covariance). Enfin $p(\boldsymbol{W},(\boldsymbol{\Sigma}_{i,j},\boldsymbol{\mu}_{i,j}))$ est une densité de gaussiennes classique :

$$p\left(\boldsymbol{W} \mid \left(\boldsymbol{\mu}_{i,j}, \Sigma_{i,j}\right)\right) = \frac{1}{(2\pi)^{d/2} \times \det\left(\Sigma_{i,j}\right)^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{W} - \boldsymbol{\mu}_{i,j}\right)^T \Sigma_{i,j} \left(\boldsymbol{W} - \boldsymbol{\mu}_{i,j}\right)\right) \qquad \text{(eq. 2.7)}$$

où d est la dimension du problème, soit 6 dans notre cas (les réponses des filtres).

L'algorithme d'optimisation EM consiste en une amélioration itérative des paramètres des gaussiennes afin d'améliorer la représentation des données d'apprentissage par le mélange de gaussiennes. A partir d'une initialisation quelconque, les paramètres $w_{i,j}$, $\boldsymbol{\mu}_{i,j}$ et $\boldsymbol{\Sigma}_{i,j}$ sont calculés de la façon suivante:

$$post_{i,j}(\boldsymbol{W}) = \frac{w_{i,j} \times p(\boldsymbol{W} \mid (\boldsymbol{\mu}_{i,j}, \Sigma_{i,j}))}{\sum_{t=1}^{K_i} w_{i,t} \times p(\boldsymbol{W} \mid (\boldsymbol{\mu}_{i,t}, \Sigma_{i,t}))}$$

$$w_{i,j} = \frac{\sum_{p=1}^{H_i} post_{i,j}(\boldsymbol{W}_p)}{H_i}, \quad \boldsymbol{\mu}_{i,j} = \frac{\sum_{p=1}^{H_i} \boldsymbol{W}_p \times post_{i,j}(\boldsymbol{W}_p)}{\sum_{p=1}^{H_i} post_{i,j}(\boldsymbol{W}_p)}$$

$$\Sigma_{i,j} = \frac{\sum_{p=1}^{H_i} post_{i,j}(\boldsymbol{W}_p) \times (\boldsymbol{W}_p - \boldsymbol{\mu}_{i,j}) \times (\boldsymbol{W}_p - \boldsymbol{\mu}_{i,j})^T}{\sum_{p=1}^{H_i} post_{i,j}(\boldsymbol{W}_p)} \qquad \text{(eq. 2.8)}$$

où H_i est le nombre de vecteurs de données d'apprentissage $\boldsymbol{W_p}$ ($1 \leq p \leq H_i$).

Le processus est itéré tant que la vraisemblance logarithmique du modèle est améliorée d'au moins un 1% d'une itération à l'autre. La vraisemblance logarithmique étant calculée de la façon suivante :

$$\log\_V(G_{i,j}) = \log \prod_{p=1}^{H} p(\boldsymbol{W}_p \mid (\boldsymbol{\mu}_{i,j}, \Sigma_{i,j})) \qquad \text{(eq. 2.9)}$$

Le principe MDL conduit lui à minimiser la fonction :

$$MDL(G_{i,j}) = -\log\_V(G_{i,j}) + \frac{m_{K_i}}{2} \log H$$

$$m_{K_i} = (K_i - 1) + K_i d + K_i \frac{d(d+1)}{2} \qquad \text{(eq. 2.10)}$$

où m_{K_i} est le nombre de paramètres nécessaires pour un modèle à K_i gaussiennes avec d=2 pour le premier modèle.

Avec nos données, nous avons trouvé des valeurs optimales K_1=2 et K_2=2 pour le premier modèle. Pour le second modèle, nous trouvons également K=2 gaussiennes pour décrire la distribution des descripteurs locaux.

La figure 2.15 présente la distribution des réponses des descripteurs du second modèle (pour pouvoir représenter cette distribution, on a seulement considéré les deux principaux

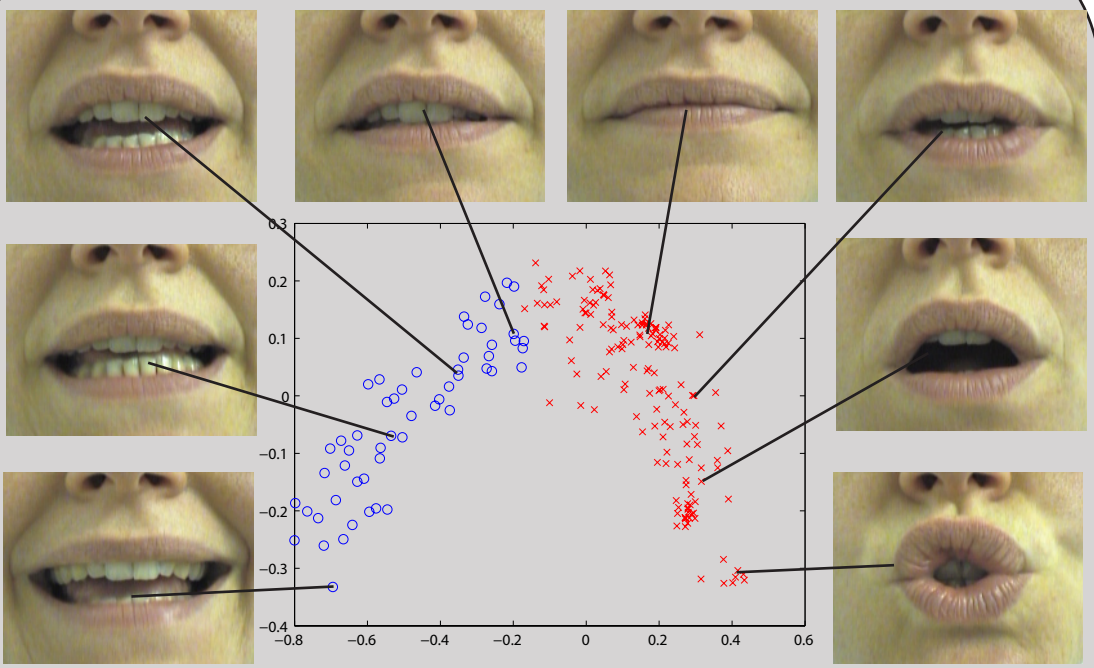


Figure 2.15 :
Distribution des données sur les deux principaux modes (80% de la variance totale) des descripteurs de commissures, avec des exemples des images d'apprentissage correspondante.
Rond bleu ou croix rouges en fonction de laquelle des deux gaussiennes (déterminées par l'algorithme EM) l'image est la plus proche.

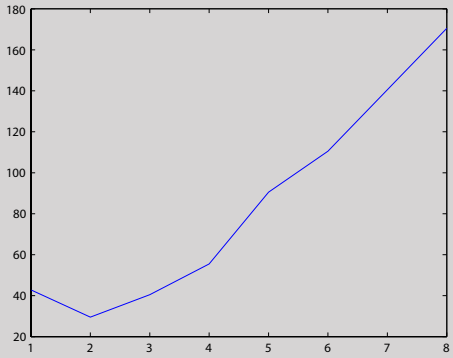


Figure 2.16 :
Valeur du critère du principe MDL en fonction du nombre de gaussiennes du modèle de commissures des lèvres.
Comme on peut le supposer intuitivement sur la figure 2.14, le nombre optimal de gaussiennes pour modéliser la distribution est égal à 2. Pour un nombre de gaussiennes supérieur, la valeur du critère augmente à peu près linéairement avec la complexité du modèle..

modes propres qui représentent 80% de la variance totale). La figure 2.16 présente l'évolution du critère MDL en fonction de K.

On constate que le nombre de gaussiennes fixé par le principe MDL correspond au nombre que l'on aurait choisi intuitivement en observant la distribution sur la figure 2.15. On observe également que quand K augmente le critère MDL augmente de façon quasi-linéaire (deux gaussiennes suffisent déjà à décrire correctement la distribution, la vraisemblance logarithmique est donc stable tandis que la complexité du modèle augmente).

Outre qu'elle présente de quelle gaussienne chaque image est la plus proche, la figure 2.15 montre également des exemples d'images le long de la distribution. On voit que l'on passe de gauche à droite de façon continue d'une bouche «étirée» à une «bouche en O» (de haut en bas on a un phénomène d'ouverture).

2.3.1.3 Détection des commissures :

Si $\boldsymbol{W}$ est le vecteur contenant les réponses des descripteurs locaux pour un point donné (36 valeurs ramenées par ACP à 6), la probabilité que ce pixel appartienne à l'une des deux catégories c_i (avec c_1 : commissure droite et c_2 : commissure gauche) sera $p(\boldsymbol{W}|y_i)$:

$$p(\boldsymbol{W} \mid y_i) = \sum_{j=1}^{K_i} w_{i,j} \times p\left(\boldsymbol{W} \mid (\boldsymbol{\mu}_{i,j}, \Sigma_{i,j})\right) \qquad \text{(eq. 2.11)}$$

Si $\boldsymbol{V}$ contient les réponses des filtres pour un couple de points (72 valeurs (36 x 2) ramenées par ACP à 6), la probabilité qu'il s'agisse des points des commissures est $p(\boldsymbol{V})$ (similaire à l'équation 2.11 mais avec seulement une catégorie, donc).

$$p(\boldsymbol{V}) = \sum_{j=1}^{K} w_j \times p\left(\boldsymbol{V} \mid (\boldsymbol{\mu}_j, \Sigma_j)\right) \qquad \text{(eq. 2.12)}$$

Ces deux modèles sont donc utilisés successivement pour détecter les points de commissures. Tout d'abord la réponse des filtres est calculée pour chaque point de la ligne d'intérêt (la ligne de minima de luminance) puis l'on calcule pour chaque point $p(\boldsymbol{W}|y_i)$. Ensuite sont retenus les six points ayant obtenu les meilleures probabilités, trois pour chacune des catégories y_i. Enfin $P(\boldsymbol{V})$ est calculé pour chacun des 9 couples de points possibles, le couple correspondant à la valeur la plus élevée étant retenu comme commissures. Ces différentes étapes de sélection sont illustrées sur un exemple dans la Figure 2.17.

On procède environ à une quarantaine d'évaluations pour le premier modèle, son utilité se justifiant par le fait que si on se passait de cette première étape, il faudrait alors 40^2=1600 évaluations du second modèle pour tester tous les couples possibles à priori (au lieu de 9 évaluations en retenant d'abord 6 candidats).

Cette détection de commissures donne une bonne estimation initiale de la position et de l'échelle de la bouche. Sur les images des bases d'apprentissage et de test, elle est robuste

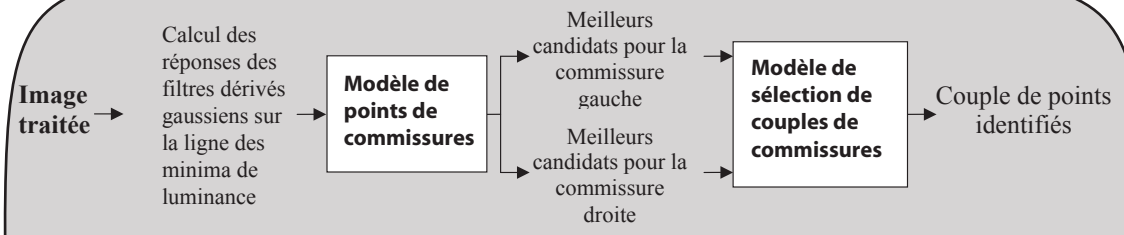
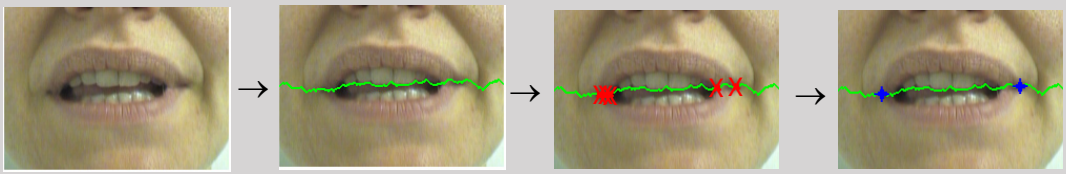


Figure 2.17 :
Détection des commissures des lèvres.
Les réponses des filtres dérivés gaussiens sont calculées sur la ligne des minima de luminance. A partir de ces réponses, un premier modèle statistique sélectionne les points les plus vraisemblables pour chaque commissure puis un second modèle désigne le couple de points le plus vraisemblable.



De gauche à droite: image traitée, ligne des minima de luminance (en vert), meilleurs candidats pour chaque commissures (croix rouges), couple réponse sélectionné (+ bleus).

dans 95% des cas pour une précision de moins de 5 % de la largeur de la bouche.

2.3.2 Modèle actif de forme et d'apparence échantillonnée

Dans cette partie nous construisons un modèle actif de forme et d'apparence en reprenant le principe de la méthodologie proposée par Cootes et largement reprise depuis pour ce type de problématique.

2.3.2.1 Echantillonnage de l'apparence

Après avoir annoté la base d'apprentissage (voir 2.2), nous avons obtenu les vecteurs $\boldsymbol{X}_i$ ($1 \leq i \leq N$) qui contiennent les coordonnées des points de contrôle définissant la forme. A partir de ces données, nous allons extraire ce que nous avons baptisé l'apparence "échantillonnée" sur chaque image. Apparence, car cela correspond aux valeurs des composantes YCbCr, et échantillonnée car ces valeurs sont extraites en 728 pixels dont les coordonnées sont obtenues par une grille d'échantillonnage elle-même déduite des $\boldsymbol{X}_i$ (figure 2.18-2) selon une méthode présentée par la suite.

La figure 2.18-1 montre que la zone bouche peut être séparée en un ensemble de quadrilatères et de triangles dont les sommets sont les points de contrôle. A partir de ce découpage, la grille est construite en rajoutant des points sur les arêtes et à l'intérieur des polygones.

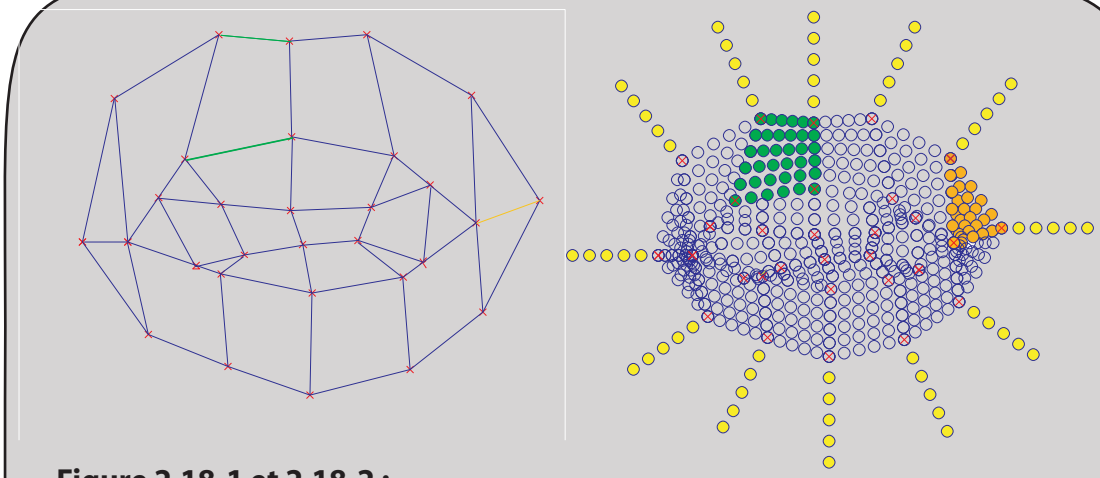


Figure 2.18-1 et 2.18-2 :
Construction de la grille utilisée pour l'échantillonnage de l'apparence.
Points de contrôle en rouge et arêtes des polygones utilisés pour construire la grille. Echantillons de la bouche en bleu (sauf exemple d'un quadrilatère en vert et d'un triangle en orange) et de la peau en jaune.

Si l'on prend l'exemple d'un quadrilatère, des points sont rajoutés sur les arêtes vertes (figure 2.18-1) par interpolation linéaire puis d'autres points sont interpolés pour relier les points des arêtes opposées pour "remplir" les polygones (ce qui donne les points verts de la figure 2.18-2). Pour un exemple de triangle, des points sont rajoutés sur l'arête orange (figure 2.18-1) par interpolation linéaire puis d'autres points sont interpolés pour relier les points de l'arête au sommet opposé (ce qui donne les points oranges de la figure 2.18-2).

Pour les polygones situés dans l'épaisseur des lèvres, on rajoute 4 points par arête, pour les polygones situés sur les dents et dans le fond de la bouche on rajoute 2 points par arête.

Enfin les échantillons de peau (points jaunes sur la figure 2.18-2) sont prélevés à partir des points de contrôle dans des directions prédéterminées (angle de 0 pour la commissure droite puis on rajoute $\pi/6$ à chaque point de contrôle en parcourant le contour dans le sens anti-horaire).

Ainsi, lorsque les points de contrôle bougent, la grille sera modifiée en fonction et chaque échantillon YCbCr sera parfaitement défini et situé. En outre, la grille se déduisant des contours de façon linéaire, le passage entre les coordonnées des points de contrôle et des points "pixels" se fait par un simple calcul matriciel.

Le fait d'utiliser cette grille d'échantillonnage définit donc précisément si un échantillon d'apparence appartient à la peau, aux lèvres, aux dents ou au fond de la bouche. Cela est particulièrement adapté pour la zone de l'intérieur de la bouche qui a un comportement non-linéaire : la bouche peut-être ouverte ou fermée, les dents (ou la langue) peuvent être apparentes ou non.

L'intérieur de la bouche est souvent modélisé de façon floue du fait que les valeurs des pixels ne sont pas suffisamment bien localisées. Ici un échantillon correspondant aux dents ne pourra jamais se retrouver avec une valeur proche du fond de la bouche (dans le

cas où les dents n'existent pas, l'échantillon existera toujours mais sera confondu avec le contour intérieur).

Les valeurs YCbCr correspondant à l'apparence sont alors sauvegardées dans les vecteurs de 2184 valeurs (728x3) notés $\boldsymbol{A}_i$ ($1 \leq i \leq N$).

En outre, il est à noter que contrairement à la méthode de prélèvement des niveaux de gris utilisée dans [Cootes, 1998] où les visages devaient être déformés par triangulation pour ramener les points de contrôle sur la forme moyenne, l'emploi de cette grille ne nécessite pas de telle opération, l'échantillonnage des valeurs étant par construction normalisé par rapport à la forme de chaque bouche.

Enfin, après une segmentation des lèvres, les échantillons de l'apparence peuvent être interpolés afin d'obtenir une bouche synthétique fidèle à l'image initiale, comme cela sera mis en oeuvre au chapitre 4.

2.3.2.2 Analyses en composantes principales

Nous allons procéder à présent à deux Analyses en Composante Principale (ACP) indépendantes l'une de l'autre sur les vecteurs de données $\boldsymbol{X}_i$, (forme) et $\boldsymbol{A}_i$ (apparence) qui constituent donc notre vérité terrain. Pour cela les valeurs moyennes ($\overline{\boldsymbol{X}}$ et $\overline{\boldsymbol{A}}$), les matrices de covariances ($\mathbf{Cov_x}$ et $\mathbf{Cov_a}$) sont calculées :

$$\overline{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i \;;\; \mathbf{Cov_s} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^T$$
$$\overline{\boldsymbol{A}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{A}_i \;;\; \mathbf{Cov_a} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{A}_i - \overline{\boldsymbol{A}}\right)\left(\boldsymbol{A}_i - \overline{\boldsymbol{A}}\right)^T \qquad \text{(eq. 2.13)}$$

On calcule ensuite les vecteurs propres ($\boldsymbol{p_{x,j}}$ et $\boldsymbol{p_{a,n}}$) ainsi que les valeurs propres correspondantes ($\lambda_{x,j}$ et $\lambda_{a,n}$) des matrices de covariance, avec $1 \leq j \leq 60$ (30 points et 2 coordonnées) et $1 \leq n \leq 2184$. A noter que les vecteurs et valeurs propres sont ordonnés dans l'ordre décroissant des valeurs propres, la valeur de ces dernières indiquant la part de la variance totale des données représentée par le mode propre correspondant.

Les vecteurs propres des matrices de covariance correspondent donc à des modes de variations des données. Enfin, comme les vecteurs propres associés aux valeurs propres les plus élevés décrivent la part la plus significative de la variance totale des données, une réduction de dimension peut être effectuée en ne conservant que les vecteurs propres les plus importants.

Nous avons décidé de garder 95% de la variance totale pour la forme et 90% pour l'apparence, les nombres nb_x et nb_a de vecteurs propres conservés dans chaque cas sont donc définis tels que :

$$\frac{\sum_{j=1}^{nb_x}\lambda_{x,j}}{\sum_{j=1}^{60}\lambda_{x,j}} \geq 95\% \geq \frac{\sum_{j=1}^{nb_x-1}\lambda_{x,j}}{\sum_{j=1}^{60}\lambda_{x,j}} \;,\; \frac{\sum_{j=1}^{nb_a}\lambda_{a,j}}{\sum_{j=1}^{2184}\lambda_{a,j}} \geq 90\% \geq \frac{\sum_{j=1}^{nb_a-1}\lambda_{a,j}}{\sum_{j=1}^{2184}\lambda_{a,j}} \qquad \text{(eq. 2.14)}$$

Avec nos données, ces seuils sélectionnent les nb_x=6 premiers vecteurs propres de variation de la forme des lèvres et les nb_a=18 premiers modes de l'apparence. Ainsi la forme peut, par exemple, être décrite par un jeu limité de 6 paramètres correspondant au poids associé à chacun des modes propres conservés.

Une grande partie des 5% de variance supprimés pour décrire la forme correspond au bruit de l'étiquetage (un même opérateur annotant deux fois de suite la même image ne le faisant jamais exactement de la même manière). De même, l'information d'apparence liée aux valeurs YCbCr des pixels est fortement sensible au bruit (l'application d'un filtre moyenneur en prétraitement permet de diminuer très sensiblement le nombre de modes propres). Sélectionner un taux de variance inférieur à celui de la forme permet de réduire de façon très importante le nombre de modes propres et de ne conserver que ceux ayant un intérêt réel.

Les vecteurs propres correspondant à la variance conservée sont alors placés dans les matrices $\mathbf{P}_x$ (forme) et $\mathbf{P}_a$ (apparence) :

$$\mathbf{P}_x = \left[\boldsymbol{p}_{x,1} \cdots \boldsymbol{p}_{x,j} \cdots \boldsymbol{p}_{x,nb_x} \right], \mathbf{P}_a = \left[\boldsymbol{p}_{a,1} \cdots \boldsymbol{p}_{a,j} \cdots \boldsymbol{p}_{a,nb_a} \right] \quad \text{(eq. 2.15)}$$

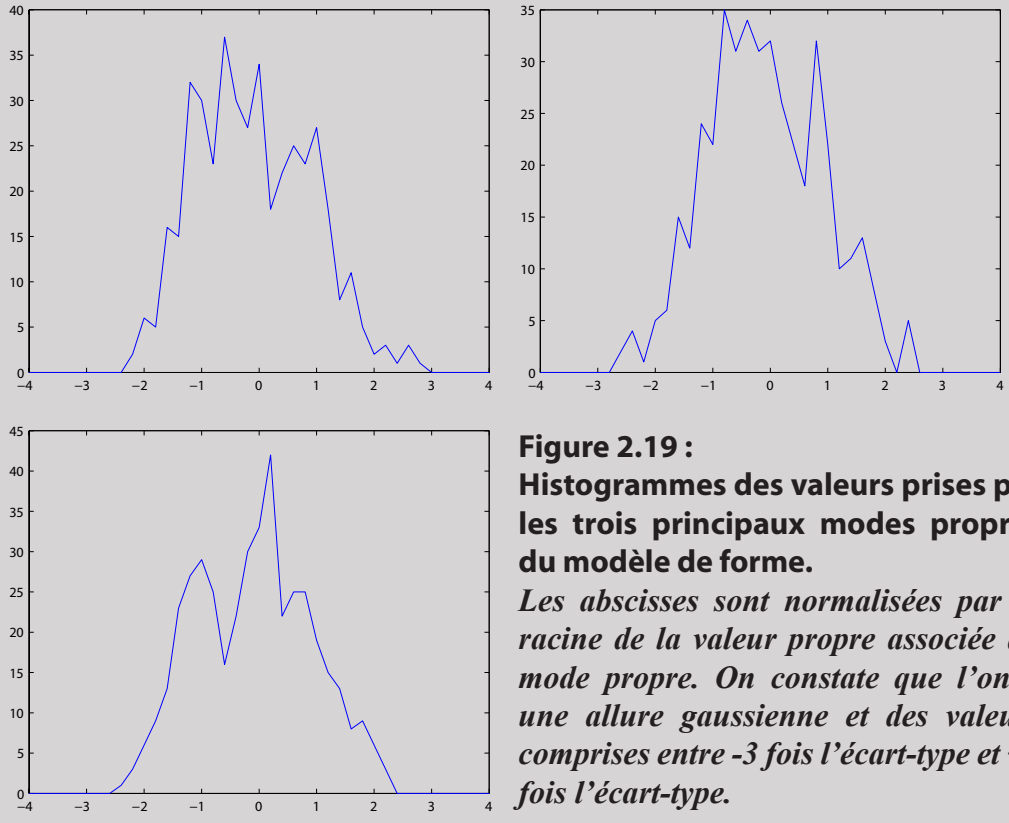


Figure 2.19 :
Histogrammes des valeurs prises par les trois principaux modes propres du modèle de forme.
Les abscisses sont normalisées par la racine de la valeur propre associée au mode propre. On constate que l'on a une allure gaussienne et des valeurs comprises entre -3 fois l'écart-type et +3 fois l'écart-type.

Toute forme et apparence de la base d'apprentissage peut donc être reproduite (à la variance perdue lors de la sélection des modes près) et de nouvelles formes plausibles peuvent être synthétisées en ajustant les vecteurs de paramètres de poids $\boldsymbol{x}$ et $\boldsymbol{a}$ dans les équations contrôlant les forme $\boldsymbol{X}$ et apparence $\boldsymbol{A}$ labiales :

$$\begin{aligned} \boldsymbol{X} &= \overline{\boldsymbol{X}} + \mathbf{P}_x \boldsymbol{x} \\ \boldsymbol{A} &= \overline{\boldsymbol{A}} + \mathbf{P}_a \boldsymbol{a} \end{aligned} \qquad \text{(eq. 2.16)}$$

Par construction, les vecteurs $\boldsymbol{x}$ et $\boldsymbol{a}$ ont des valeurs moyennes nulles et sont supposés respecter une loi gaussienne. Leurs valeurs seront donc dans 95% comprises dans un intervalle équivalent à deux fois l'écart type, et dans 99% dans un intervalle de trois fois l'écart type. Les relations suivantes sont donc vérifiées dans 99% des cas :

$$\begin{aligned} -3\sqrt{\lambda_{x,j}} \le x(j) \le 3\sqrt{\lambda_{x,j}} \text{ pour } 1 \le j \le n \\ -3\sqrt{\lambda_{a,k}} \le a(k) \le 3\sqrt{\lambda_{a,k}} \text{ pour } 1 \le k \le n \end{aligned} \qquad \text{(eq. 2.17)}$$

La figure 2.19 présente la distribution des poids des trois principaux modes propres du modèle de forme. On constate que l'allure est bien gaussienne et que les limites de l'équation 2.17 sont valides.

La figure 2.20 montre, quant à elle, les effets des six premiers modes de variation de la forme. Les valeurs du vecteur de poids $\boldsymbol{x}$ varient de -2 fois l'écart-type à +2 fois l'écart-type. On observe que le premier mode correspond à une ouverture et le deuxième à une protusion. Le quatrième mode contrôle en partie le mouvement des dents, les autres modes étant plus ardus à interpréter.

L'étape suivante consiste à relier entre eux les modèles de forme et d'apparence d'une façon similaire à [Cootes, 1998].

L'approche est de procéder à une nouvelle ACP, dite de second niveau, avec les valeurs des paramètres de poids des deux modèles afin d'obtenir un modèle statistique qui lie les variations de forme (représentée par les $\boldsymbol{x}_i$) et les variations d'apparence (représentée par les $\boldsymbol{a}_i$). Nous aurons alors une modélisation conjointe et cohérente des deux grandeurs qui sont évidemment fortement liées l'une à l'autre. En effet, pour prendre l'exemple de la zone de l'intérieur de la bouche, les échantillons des dents et du fond de la bouche prennent des valeurs très différentes selon que la bouche est ouverte ou non et que le contour des dents est distinct du contour intérieur de la bouche ou non.

Afin de pouvoir lier les variations des modes propres de chaque modèle, nous devons tout d'abord calculer pour chaque image de la base d'apprentissage les valeurs correspondantes des vecteurs de poids $\boldsymbol{x}_i$ et $\boldsymbol{a}_i$ avec $1 \le i \le N$ en inversant les équations 2.16, les données réelles $\boldsymbol{X}_i$ et $\boldsymbol{A}_i$ étant parfaitement connues :

$$\boldsymbol{x}_i = \mathbf{P}_x^T \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}} \right) ; \boldsymbol{a}_i = \mathbf{P}_a^T \left(\boldsymbol{A}_i - \overline{\boldsymbol{A}} \right) \qquad \text{(eq. 2.18)}$$

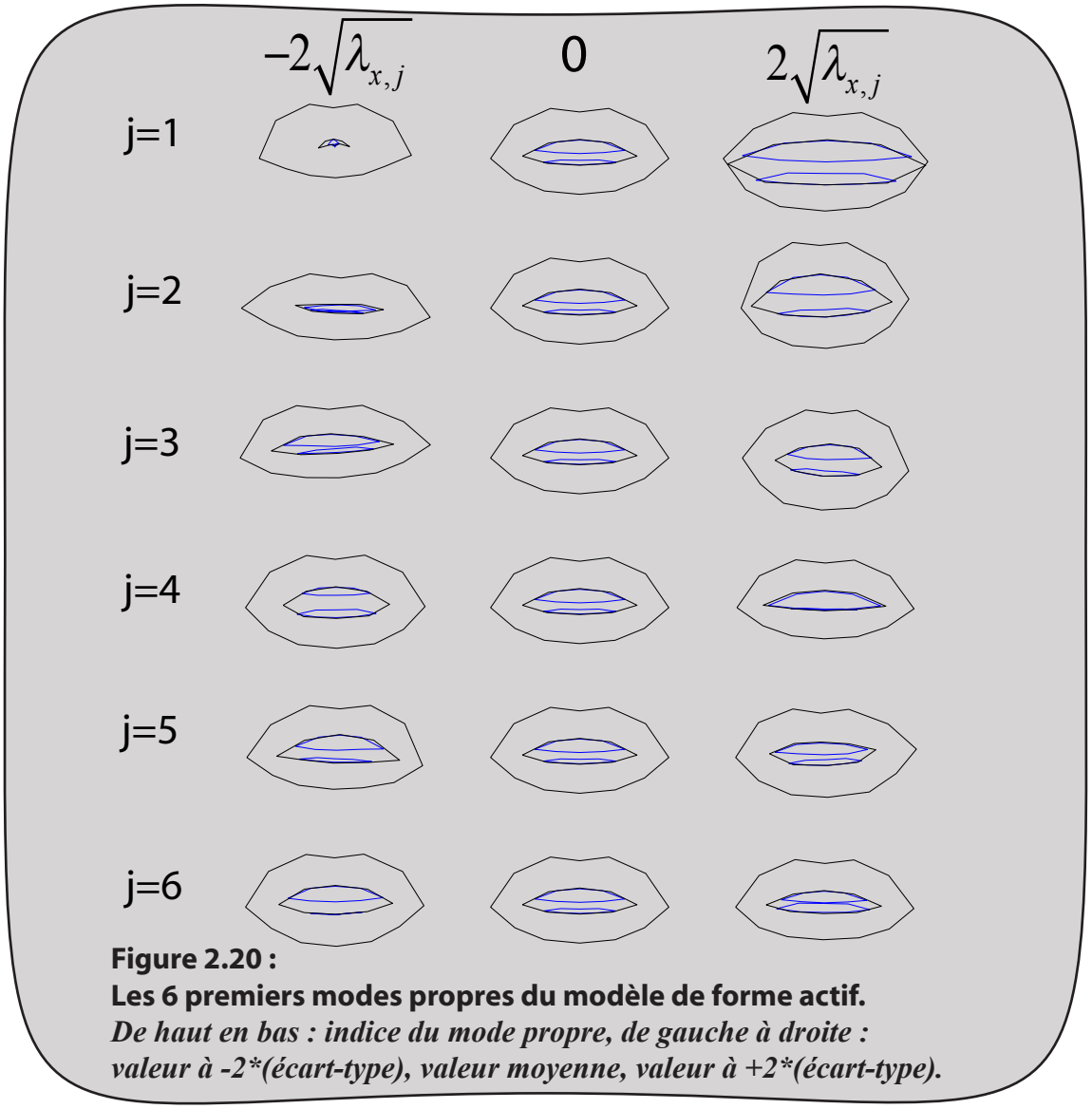


Figure 2.20 :
Les 6 premiers modes propres du modèle de forme actif.
De haut en bas : indice du mode propre, de gauche à droite : valeur à -2*(écart-type), valeur moyenne, valeur à +2*(écart-type).

Avant de procéder à l'ACP, il est en outre nécessaire de normaliser les différences d'unités entre les informations de forme (coordonnées normalisées) et d'apparence (niveaux de gris). Nous souhaitons aussi augmenter le poids de la forme par rapport à l'apparence, considérant en effet que dans un cas mono-locuteur les variations d'apparence sont causées principalement par les variations de forme (ce qui va en outre diminuer le nombre de modes retenus pour le modèle combiné, la variance due à l'apparence étant diminuée). Cela est effectué par le coefficient de pondération W :

$$W = W_e \times \frac{\sum_{j=1}^{nb_a} \lambda_{a,j}}{\sum_{j=1}^{nb_x} \lambda_{x,j}} \text{ avec } W_e = 2 \qquad \text{(eq. 2.19)}$$

Le coefficient W_e=2 a été déterminé empiriquement comme étant la limite permettant de diminuer le nombre de modes de variations du modèle combiné sans pour autant dégrader les performances de la segmentation. Si l'on augmentait cette valeur, jusqu'à 100 par exemple, les modes de variations du modèle combiné seraient en fait ceux du modèle de forme et l'apparence se déduirait alors totalement de la forme. Le nombre de paramètres à optimiser serait alors minimal, mais la moindre variation d'éclairage ou d'apparence du locuteur (un rougissement de la peau, par exemple) dégraderait les performances du modèle qui ne serait plus capable de s'adapter.

Affecter un poids supérieur à la forme se rapproche dans le principe de [Cootes, 1998+], dans lequel seuls les paramètres du modèle de forme sont optimisés lors de la segmentation, les paramètres du modèle de niveau de gris étant calculés directement en fonction de l'image observée. Dans [Hou, 2001], les auteurs proposaient, quant à eux, de déduire la forme directement de l'apparence. Outre que cette approche n'est pas vraiment intuitive (il est plus évident de supposer que les mouvements des contours des lèvres changent les valeurs des pixels que l'inverse) et qu'elle conduit à considérer un nombre supérieur de modes de variations, elle repose sur une présomption de totale dépendance de la forme par rapport à l'apparence qui n'est pas robuste dans tous les cas, particulièrement dans une optique multi-locuteurs qui est l'objectif de cette thèse, sinon de ce chapitre.

Les vecteurs combinés $\boldsymbol{C}_i$ sur lesquels va être effectué l'ACP de second niveau sont donc construits de la façon suivante :

$$\boldsymbol{C}_i = \begin{bmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{a}_i \end{bmatrix}, 1 \le i \le N \qquad \text{(eq. 2.20)}$$

$\overline{\boldsymbol{C}}$ est la moyenne du vecteur combiné qui est approximativement nulle puisque les vecteurs de poids ont une distribution gaussienne avec une moyenne nulle par construction. $\mathbf{Cov_c}$ est la matrice de covariance et $\mathbf{P_c}$ est une matrice contenant les vecteurs propres représentant 98% de la variance ce qui donne nb_c=10 modes dans notre cas avec les valeurs propres λ_k, $1 \le k \le 10$. Sans augmenter le poids de la forme (c'est à dire avec W_e=1), on aurait eu nb_c=15 modes.

$$\overline{\boldsymbol{C}} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{a}_i \end{pmatrix} \simeq 0,\ \mathbf{Cov_c} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{a}_i \end{pmatrix}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{a}_i \end{pmatrix}^T \qquad \text{(eq. 2.21)}$$

Finalement toute forme et apparence échantillonnée de la base d'apprentissage (là encore aux pertes de variance près), ou de nouveaux exemples, peuvent être générés en ajustant simplement $\boldsymbol{c}$ dans l'équation suivante :

$$(1)\ \boldsymbol{C} = \begin{bmatrix} W.\boldsymbol{x} \\ \boldsymbol{a} \end{bmatrix} = \mathbf{P_c}\boldsymbol{c} \Rightarrow \begin{cases} (2)\ \boldsymbol{X} = \overline{\boldsymbol{X}} + \mathbf{P_x}\boldsymbol{x} \\ (3)\ \boldsymbol{A} = \overline{\boldsymbol{A}} + \mathbf{P_a}\boldsymbol{a} \end{cases} \qquad \text{(eq. 2.22)}$$

L'équation 2.22-2 contrôle le modèle de forme, l'équation 2.22-3 contrôle l'apparence et enfin l'équation 2.22-1 contrôle le modèle combinant la forme et l'apparence.

La figure 2.21 présente l'évolution du nombre de paramètres en fonction de la variance expliquée pour chacun des trois modèles.

Enfin, la figure 2.22 présente les six premiers modes du modèle combiné. Le premier mode est une ouverture, le second une protusion, les autres modes étant plus ardus à interpréter.

Ainsi segmenter la bouche sur une image inconnue consistera à optimiser les paramètres de ces modèles actifs sans oublier que pour passer de la forme normalisée $\boldsymbol{X}$ à la forme réelle $\boldsymbol{X_r}$, il convient également de connaître la position $\boldsymbol{t}$ et l'échelle q de la bouche (en supposant la bouche toujours horizontale) puisque :

$$\boldsymbol{X}_r = q\boldsymbol{X} + \boldsymbol{t} \qquad \text{(eq. 2.23)}$$

Donc si l'on utilise le modèle combiné forme/apparence il faudra résoudre un problème de dimension 13 (nombre de valeurs de $\boldsymbol{c}$ + 3), tandis que si l'on utilise seulement le modèle de forme le problème sera de dimension 9 (le nombre de valeurs $\boldsymbol{x}$ + 3).

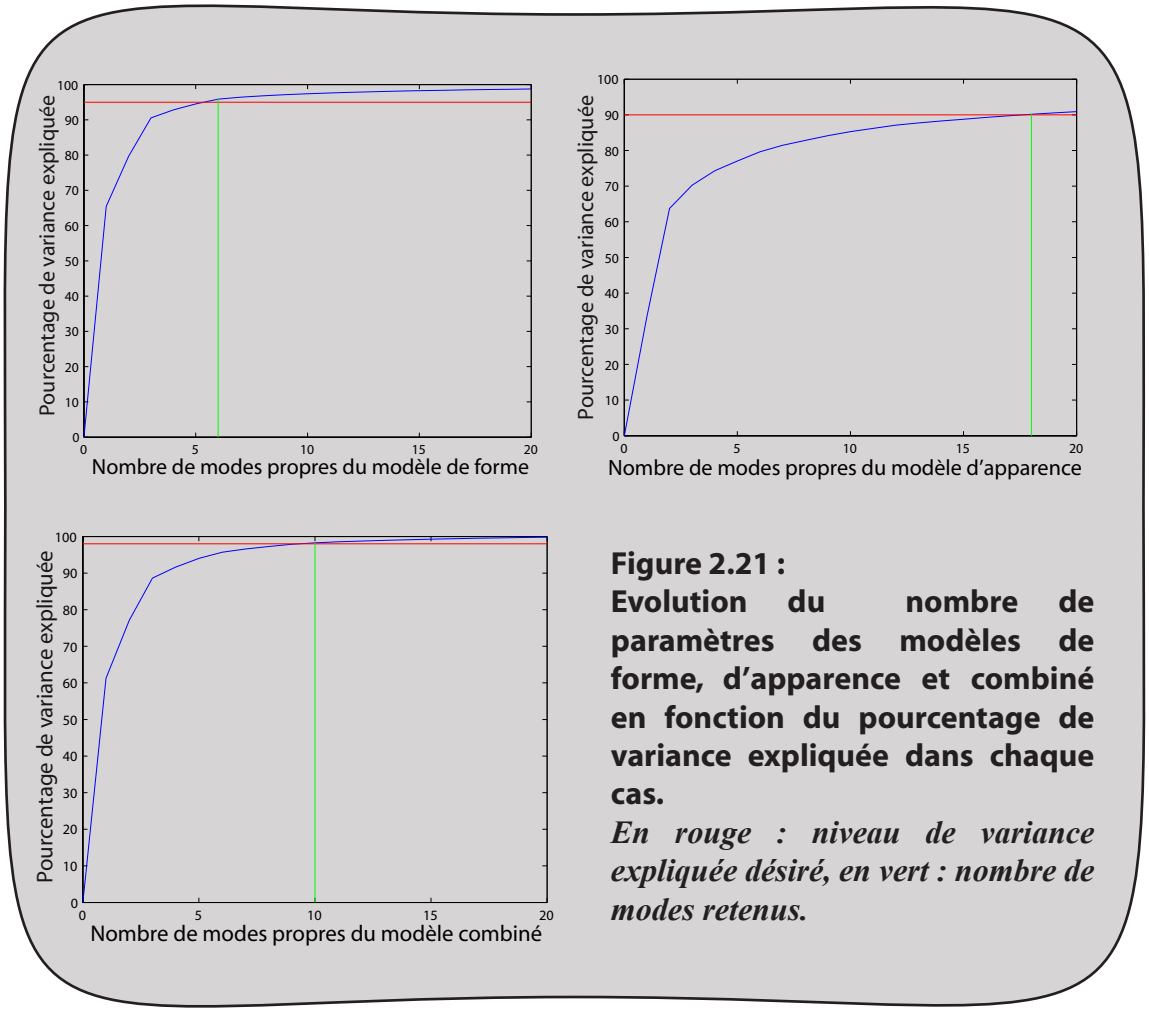


Figure 2.21 :
Evolution du nombre de paramètres des modèles de forme, d'apparence et combiné en fonction du pourcentage de variance expliquée dans chaque cas.
En rouge : niveau de variance expliquée désiré, en vert : nombre de modes retenus.

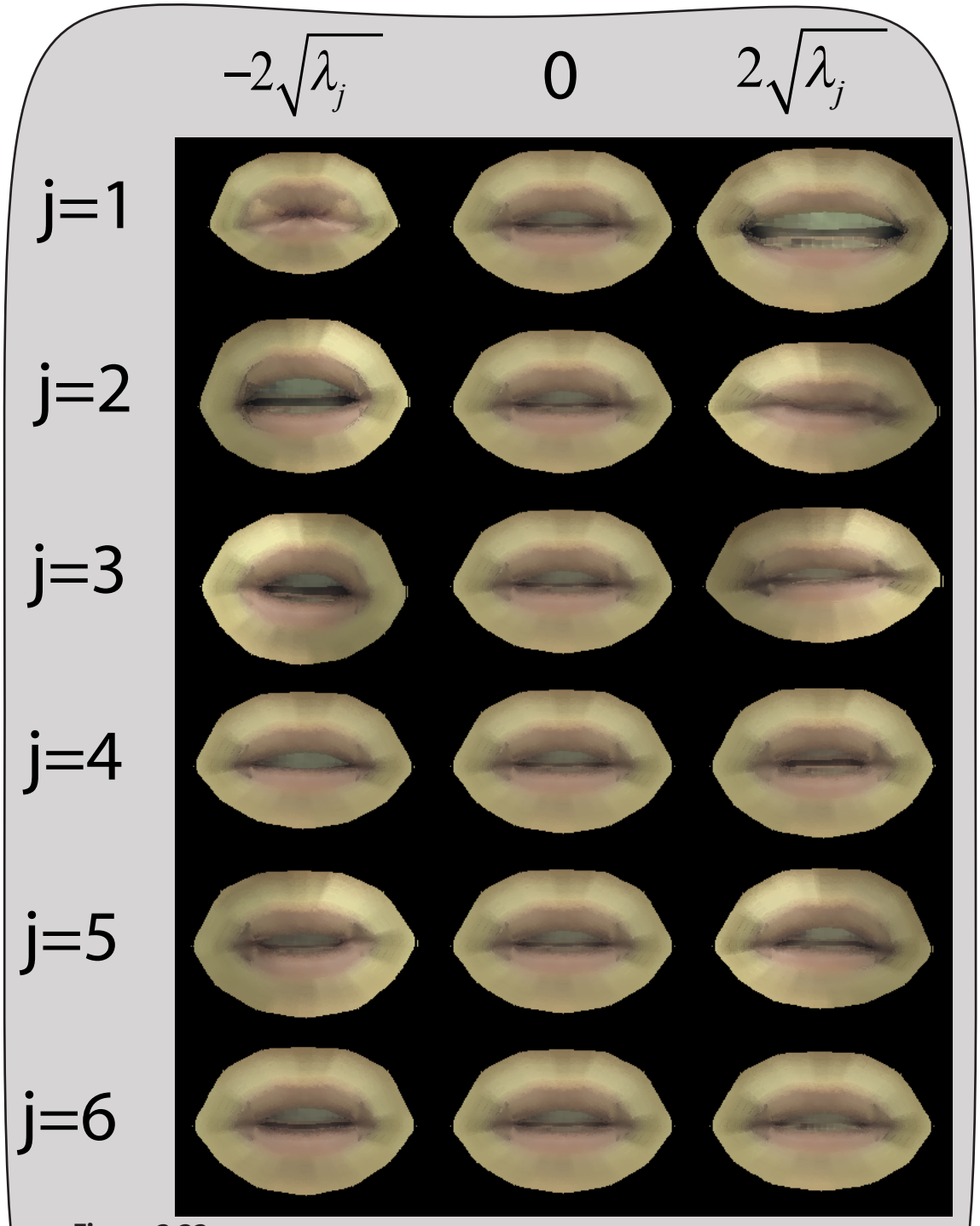


Figure 2.22 :
Les 6 premiers modes propres du modèle combiné.
L'apparence est interpolée à partir des échantillons (voir chapitre 4.1), de haut en bas : indice du mode propre, de gauche à droite : valeur à -2*(écart-type), valeur moyenne, valeur à +2*(écart-type).

2.3.2.3 Apprentissage des propriétés statistiques de chaque EGB

La dernière étape de l'apprentissage des modèles actifs est de calculer les propriétés statistiques des paramètres des modèles selon l'EGB, ce qui est une opération très facile puisque l'EGB est connu pour chaque image de la base de l'apprentissage.

Les valeurs du vecteur $\boldsymbol{c}$ sont d'abord calculées pour chaque image :

$$\boldsymbol{c}_i = \mathbf{P}_{\mathbf{c}}^{\,T} \begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{a}_i \end{pmatrix}, \textit{pour } 1 \le i \le N \qquad \text{(eq. 2.24)}$$

Ensuite les moyennes de chaque paramètre sont calculées pour chacun des EGB et sauvegardées dans les vecteurs $\boldsymbol{c}_j^{\boldsymbol{egb}}$ (avec $1 \le j \le 4$). Les vecteurs $\boldsymbol{c}_j^{\boldsymbol{b}}$ ($1 \le j \le 4$) contenant les variations maximales des paramètres par rapport à ces moyennes sur la base d'apprentissage sont également calculés pour chaque EGB. Il est à noter que l'on considérera cette amplitude comme étant au minimum égale à l'écart-type, ceci afin de ne pas trop contraindre dans la suite l'optimisation des paramètres si l'estimation de l'EGB était mauvaise.

$$\text{pour } \boldsymbol{1} \le \boldsymbol{j} \le \boldsymbol{4}, \quad \boldsymbol{c}_j^{\boldsymbol{egb}} = \frac{\mathbf{1}}{M_j} \sum_{k \in [\![\mathbf{1},\mathbf{N}]\!] / \{\mathrm{EGB}(k)=j\}} \boldsymbol{c}_k \qquad \text{avec } M_j = \underset{k \in [\![\mathbf{1},\mathbf{N}]\!]}{card} \left(\mathrm{EGB}(k) = j\right)$$

$$\text{pour } \boldsymbol{1} \le j \le \boldsymbol{4} \textbf{ et pour } \boldsymbol{1} \le i \le nb_c \textbf{ :}$$

$$\boldsymbol{c}_j^{\boldsymbol{b}}(i) = \max\left(\sqrt{\lambda_{\boldsymbol{c},j}}, \left(\max_{k \in [\![\mathbf{1},\mathbf{N}]\!] / \{\mathrm{EGB}(k)=j\}} \left(\boldsymbol{c}_k(i)\right) - \boldsymbol{c}_j^{\boldsymbol{egb}}(i)\right), \left(\boldsymbol{c}_j^{\boldsymbol{egb}}(i) - \min_{\mathbf{k} \in [\![\mathbf{1},\mathbf{N}]\!] / \{\mathrm{EGB}(k)=j\}} \left(\boldsymbol{c}_k(i)\right)\right)\right) \qquad \text{(eq. 2.25)}$$

On obtient de façon parfaitement analogue $\boldsymbol{x}_j^{\boldsymbol{egb}}$ et $\boldsymbol{x}_j^{\boldsymbol{b}}$ ($1 \le j \le 4$) contenant les moyennes et variations maximales permises des paramètres du modèle de forme pour chaque EGB.

On peut également remarquer que les EGB 3 et 4 représentant le sourire et la protrusion correspondent (en observant la figure 2.20) aux variations extrêmes des deux premiers modes propres du modèle de forme.

2.4 Segmentation d'une image inconnue

2.4.1 Principe Général

Nous avons construit lors de la partie précédente un modèle actif qui peut générer potentiellement n'importe quelle forme de bouche en réglant un nombre limité de paramètres (la limite étant que la configuration souhaitée doit avoir été observée dans la base d'apprentissage).

Si l'on veut segmenter le contour des lèvres sur une image inconnue, il va donc falloir être capable de régler automatiquement les paramètres des modèles actifs à des valeurs qui

modéliseront le mieux possible la bouche observée, ainsi que de déterminer la position $\boldsymbol{t}$ et le facteur d'échelle q de la bouche.

La façon la plus intuitive d'optimiser ces valeurs est de se servir d'une fonction de coût dont la réponse sera faible lorsque les paramètres modéliseront correctement la bouche et élevée quand la segmentation sera mauvaise.

Nous allons donc présenter une stratégie d'optimisation, puis tester différentes fonctions de coût pour choisir la plus adaptée au problème.

2.4.2 Méthode d'optimisation : la descente du Simplex

Nous avons donc une fonction de plusieurs variables à minimiser. Parmi les différentes méthodologies d'optimisation possible, nous avons arrêté notre choix sur la méthode de descente du simplex (Downhill Simplex Method, DSM). Cette méthode classique de minimisation/maximisation est due originellement à Nelder et Mead ([Nelder, 1965]).

L'un des principaux avantages de cette méthode (par rapport par exemple à la descente du gradient classique) est qu'elle ne nécessite pas de calculer ou d'estimer la dérivée de la fonction à optimiser. En outre, sa méthode de recherche du minimum local grâce à des transformations géométriques lui permet généralement de sortir des puits de la fonction de coût et de converger vers le minimum global même quand la fonction est très peu convexe. En revanche, elle nécessite un nombre important d'évaluations de la fonction, et est à éviter dans le cas de problèmes à très grandes dimensions. Dans le cadre de notre travail, nous avons utilisé quelques modifications permettant d'adapter le DSM au nombre de dimensions requis et une phase d'initialisation permettant de se rapprocher de la solution finale avant même le début de l'algorithme (grâce au détecteur de commissure et au test de l'EGB), ce qui diminue à la fois le nombre d'itérations et le risque de convergence vers des minima locaux. Au final, cette méthode converge la plupart du temps en relativement peu d'itérations vers le minimum absolu, tout en évitant généralement de tomber sur des minima locaux.

Par rapport aux méthodes de convergence des modèles actifs présentées au 1.2.3, notre choix est un compromis de simplicité et de précision. Par rapport à la méthode originale vue dans [Cootes, 1998] employant une matrice précalculée pour calculer et mettre à jour les paramètres, le DSM est plus à même de gérer la non convexité et la grande dimension de la fonction de coût. En effet, l'approche proposée dans [Cootes, 1998] se traduit par une descente du gradient avec une matrice de Jacobi constante et converge donc vers le premier minimum local rencontré.

Dans le principe, notre approche est voisine de celle présentée dans [Dornaika, 2004], où la convergence se déroule en trois phases dont les deux premières permettent de fournir une estimation de la solution grâce à une première convergence d'une descente de gradient et une diffusion stochastique. Dans notre travail, nous recherchons également à obtenir la meilleure estimation possible du résultat grâce à des prétraitements adaptés.

Enfin, par rapport à des méthodes plus complexes de fusion de critères (du type [Romdhani, 2005]) adaptées au cas 3D, nous avons fait le choix de n'avoir recours qu'à des techniques relativement simples et peu coûteuses.

2.4.2.1 Principe

Cette méthode a une interprétation géométrique qui en rend facile la compréhension. Si le problème à résoudre est fonction de M variables, le simplex sera une figure géométrique à M+1 sommets correspondant aux points où sera estimée la fonction à minimiser. Par exemple dans le cas d'une fonction à deux variables, le simplex sera un triangle dans l'espace des paramètres. Dans le cas d'une fonction à trois variables, nous aurons un tétrahédon.

Si on note $\boldsymbol{e}_0$ un point initial du Simplex choisit de façon pertinente, les autres M points $\boldsymbol{e}_i$ sont obtenus de la façon suivante:

$$\boldsymbol{e}_i = \boldsymbol{e}_0 + l_i \boldsymbol{n}_i,\ 1 \le i \le M \qquad \text{(eq. 2.26)}$$

où les $\boldsymbol{n}_i$ sont des vecteurs définissant une base orthonormée dans l'espace des paramètres et les l_i sont la longueur définissant l'intervalle de recherche du minimum dans chaque direction.

Dans notre cas, un choix simple serait de prendre le vecteur moyen (c'est à dire le vecteur nul) pour $\boldsymbol{e}_0$, les vecteurs propres retenus pour les $\boldsymbol{e}_i$ et 3 fois les écart-types correspondant pour les l_i. Nous verrons néanmoins, par la suite, que ces choix peuvent être faits de façon plus pertinente afin de rapprocher autant que possible le Simplex de sa valeur optimale et ce, avant même le début de la convergence.

Une fois le Simplex initial déterminé, la fonction est calculée en chacun de ces points et l'on repère le point qui correspond au minimum (qui est donc la première estimation des paramètres minimisant le problème), celui qui correspond au maximum et enfin celui qui correspond au deuxième maximum. Le Simplex va ensuite être modifié par diverses transformations géométriques afin de trouver un point qui diminuera la valeur du maximum du Simplex. Les transformations du Simplex sont présentées dans l'Annexe 2.

2.4.2.2 Arrêt

Il est difficile de choisir un critère d'arrêt pour la convergence du Simplex puisque le minimum n'est pas nécessairement amélioré à chaque itération de l'algorithme. Nous avons opté pour un double critère : il faut à la fois que l'écart entre la valeur du minimum et du maximum soit en dessous d'un certain seuil et que la dernière amélioration du minimum soit également en dessous d'un seuil fixé. En outre, et par sécurité, un nombre maximum d'itérations est fixé.

2.4.2.3 Application à la segmentation de la bouche

Nous voulons obtenir le meilleur jeu de paramètres afin que notre modèle opère la meilleure segmentation possible de la bouche.

Dans la suite nous noterons $C(I_n)(\boldsymbol{p},q,\boldsymbol{t})$ la valeur d'une fonction de coût appliquée à l'image I_n pour un vecteur de paramètres $\boldsymbol{p}$ et pour l'échelle q et la position $\boldsymbol{t}$. En pratique,

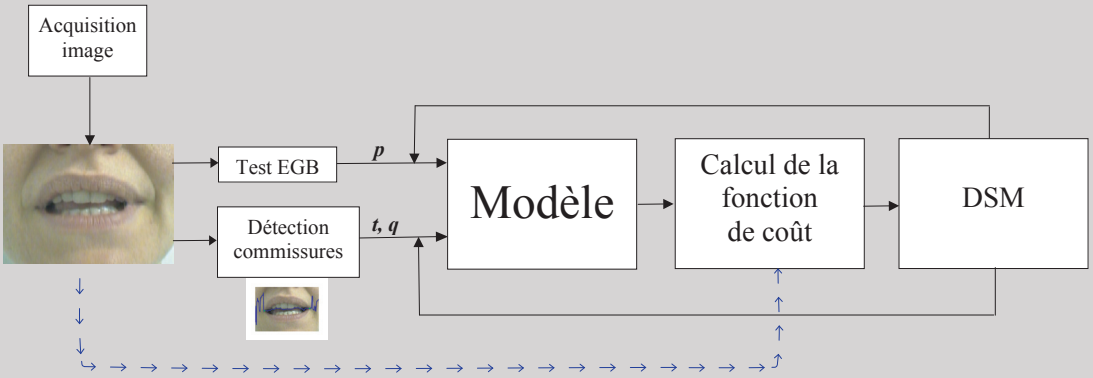


Figure 2.23 :
Schéma de principe de l'optimisation des paramètres : Initialisation des paramètres par détection des commissures et test de l'EGB puis calcul de la fonction de coût et optimisation par l'algorithme du simplex.

$\boldsymbol{p}$ sera soit $\boldsymbol{c}$ (modèle combiné forme/apparence) soit $\boldsymbol{s}$ (modèle de forme uniquement).

Un choix classique pour initialiser l'algorithme de DSM serait de prendre le vecteur nul pour $\boldsymbol{p}$ et des valeurs q et $\boldsymbol{t}$ (telles que $\boldsymbol{t}$ soient les coordonnées d'un point de la ligne de minima de luminance et q environ un tiers de la largeur de la zone du visage) pour obtenir le point initial $\boldsymbol{e}_0$ de l'équation 2.26. Les longueurs de recherche l_i (équation 2.26) seraient 3 fois l'écart-type pour les paramètres de $\boldsymbol{p}$ et un vingtième de la largeur de la zone visage pour q et $\boldsymbol{t}$.

Afin d'améliorer la précision, la robustesse ainsi que la vitesse de convergence en diminuant le risque d'obtenir un minimum local, nous allons définir la meilleure initialisation possible pour l'algorithme ainsi que des intervalles de recherche appropriés. Pour la première image, cela sera accompli en utilisant les EGB et la détection de commissures. Pour les images suivantes, on se servira des images précédentes pour effectuer un suivi des paramètres. Cette stratégie d'optimisation est présentée par la figure 2.23.

Première image I_1 d'une séquence

Le modèle de commissure est utilisé afin de détecter les deux points clés. Cela donne des valeurs initiales généralement fiables pour la position de la bouche $\boldsymbol{t}$ ainsi que pour son échelle q.

Afin d'obtenir une première estimation de $\boldsymbol{p}$, nous testons l'Etat Général de la Bouche (EGB). Pour cela, on calcule pour chaque EGB: $C(I_1)(\boldsymbol{p}_j^{egb},q,\boldsymbol{t})$ pour $1 \leq j \leq 4$, les $\boldsymbol{p}_j^{egb}$ étant les jeux de paramètres moyens calculés lors de l'apprentissage. On obtient l'indice correspondant au minimum jm, ce qui donne l'initialisation du DSM : $(\boldsymbol{p}_{jm}^{egb},q,\boldsymbol{t})$.

Les intervalles de recherche pour le vecteur de paramètres $\boldsymbol{p}$ sont définis par l'EGB, les autres points du Simplex étant déduits du point initial par les valeurs de $\boldsymbol{x}_{jm}^{b}$ telles que

calculées lors de l'apprentissage (2.3.2.3).

Nous procédons ensuite à la minimisation de $C(I_1)(\boldsymbol{p},q,\boldsymbol{t})$ par DSM et nous trouvons le jeu de paramètre optimal pour l'image $\mathbf{I}_1$: $(\boldsymbol{p}_1,q_1,\boldsymbol{t}_1)$.

Le DSM est considéré avoir convergé quand la différence entre les valeurs maximum et minimum du Simplex est en dessous d'un seuil et que la valeur du minimum est également en dessous d'un autre seuil fixé empiriquement.

Suivi de mouvement

Nous calculons tout d'abord un critère afin de déterminer si la nouvelle image a beaucoup changé par rapport à la précédente. Pour l'image I_{n+1}, il s'écrit :

$$\left|\frac{C(\mathbf{I}_{n+1})(\boldsymbol{p}_n,q_n,\boldsymbol{t}_n)-C(\mathbf{I}_{n+1})(\boldsymbol{p}_n,q_n,\boldsymbol{t}_n)}{C(\mathbf{I}_{n+1})(\boldsymbol{p}_n,q_n,\boldsymbol{t}_n)}\right| \leq 20\% \qquad \text{(eq. 2.27)}$$

Si ce critère est vérifié, cela signifie que la bouche a peu bougé par rapport à l'image précédente. Nous minimiserons donc $C(I_{n+1})(\boldsymbol{p}_n,q,\boldsymbol{t})$ par DSM ($\boldsymbol{p}_n$ étant donc l'estimation initiale de $\boldsymbol{p}$) avec des intervalles de recherche réduits correspondant à $0{,}5\sqrt{\lambda_k}$ pour le vecteur $\boldsymbol{p}$. q_n et $\boldsymbol{t}_n$ sont les estimations initiales de q et $\boldsymbol{t}$.

Si le critère est non vérifié, nous testons à nouveau les EGB et détectons à nouveau les commissures.

2.4.3 Fonctions de coût

Nous allons à présent présenter trois fonctions de coût permettant d'optimiser les valeurs des paramètres du modèle. Selon la fonction de coût, le modèle considéré sera soit le modèle de forme, soit le modèle combiné forme/apparence.

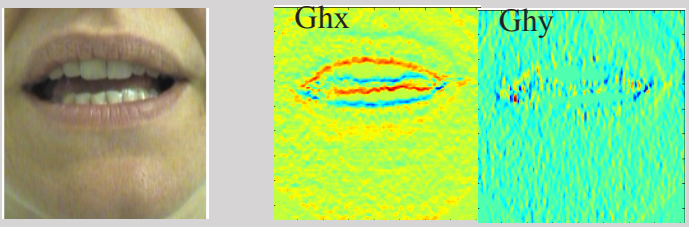
2.4.3.1 Fonction de coût basée sur un calcul de champ de gradients

Les 30 points de contrôle qui décrivent la forme peuvent être séparés en 6 courbes (figure 2.7). Si le flux d'un vecteur gradient approprié à travers ces courbes est maximisé alors les courbes colleront au plus près des contours présents dans l'image. Si $\mathbf{G}$ est un champ de vecteur gradient et si ζ est une courbe, le flux du vecteur $\mathbf{G}$ à travers ζ est :

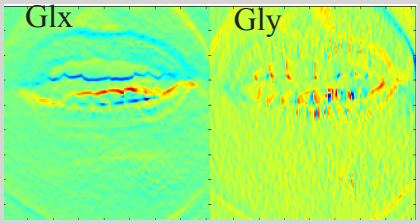
$$\mathsf{f} = \frac{\int_{\zeta} \mathbf{G}.\mathrm{d}\boldsymbol{n}}{\int_{\zeta} \mathrm{d}s} \qquad \text{(eq. 2.28)}$$

Des champs de gradients différents sont utilisés selon les courbes.

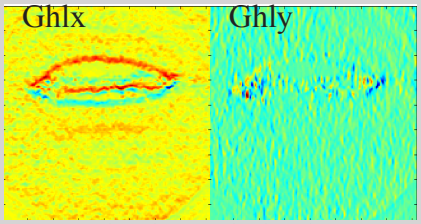
Nous avons d'abord repris le gradient hybride **Ghl** tel qu'introduit dans [Eveno, 2004]. Ce champ de gradient combine une pseudo-teinte (voir 2.1.2) normalisée $\mathbf{H_n}$ avec



Gradients de teinte.



Gradients de luminance.



Gradients hybrides.

Figure 2.24 :
Champs de vecteurs de gradient utilisés (masques de Prewitt) pour la fonction de coût C_f.
Gradients horizontaux notés Gx, gradients verticaux notés Gy.

une luminance normalisée $\mathbf{L_n}$ pour améliorer la netteté de la frontière haute de la lèvre supérieure :

$$Ghl(x,y) = \nabla\left[Hn(x,y) - Ln(x,y)\right] \qquad \text{(eq. 2.29)}$$

Les autres champs de gradient utilisés sont **Gh** et **Gl** respectivement basés sur la pseudo-teinte et la luminance.

En se reportant à la figure 2.7, **Ghl** est utilisé pour calculer le flux à travers la courbe 1. **Gh** est utilisé pour les courbes 2, 3 et 4 et **Gl** pour les courbes 5 et 6. La figure 2.24 présente les différents champs de gradients utilisés pour ce critère.

La fonction de coût définie ici et nommée C_g est donc calculée comme étant la somme de ces 6 flux à travers les contours de la forme. Il est à noter que la modélisation de l'apparence n'intervient pas, ce n'est donc qu'un critère purement orienté forme (même si le critère utilise l'apparence du voisinage immédiat du contour à travers les gradients). L'absence de prise en compte directe de l'apparence peut faire que les contours se retrouvent mal placés dans certains cas de figure (le contour entre les dents et le fond de la bouche peut être confondu avec le contour entre les lèvres et les dents par exemple) et que la segmentation précise d'une bouche fermée soit difficile (les contours de la zone intérieure correspondant alors à des gradients faibles).

La figure 2.25 présente le principe de cette fonction de coût qui n'optimise donc que le modèle de forme.

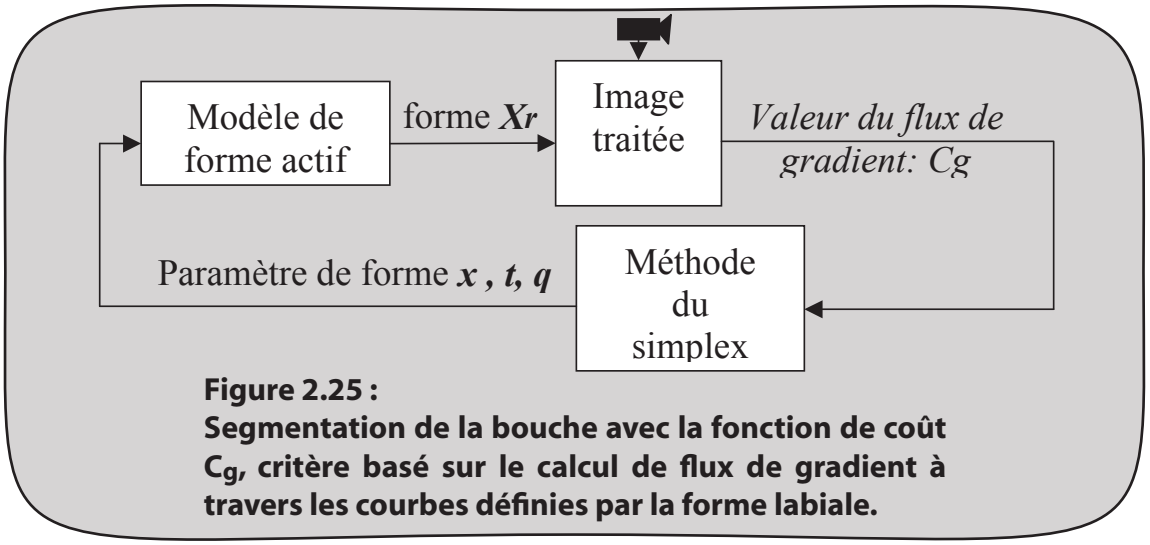


Figure 2.25 :
Segmentation de la bouche avec la fonction de coût C_g, critère basé sur le calcul de flux de gradient à travers les courbes définies par la forme labiale.

2.4.3.2 Fonction de coût basée sur les valeurs des pixels

Ici, la fonction de coût C_V compare simplement les valeurs YCbCr de l'apparence échantillonnée pour un jeu de paramètres donnés aux valeurs YCbCr observées dans l'image pour la forme correspondante en calculant l'erreur quadratique moyenne entre la prévision et l'observation, critère très largement utilisé depuis les premiers AAMs.

A partir de cette fonction simple, on peut créer un critère plus complet et global $C_c=C_V/C_g$ qui combinera des critères de forme et d'apparence, les contours de la forme labiale se plaçant sur les zones de forts gradients tout en assurant une discrimination entre les contours corrects ou non par rapport au simple critère C_g (le contour dents/fond

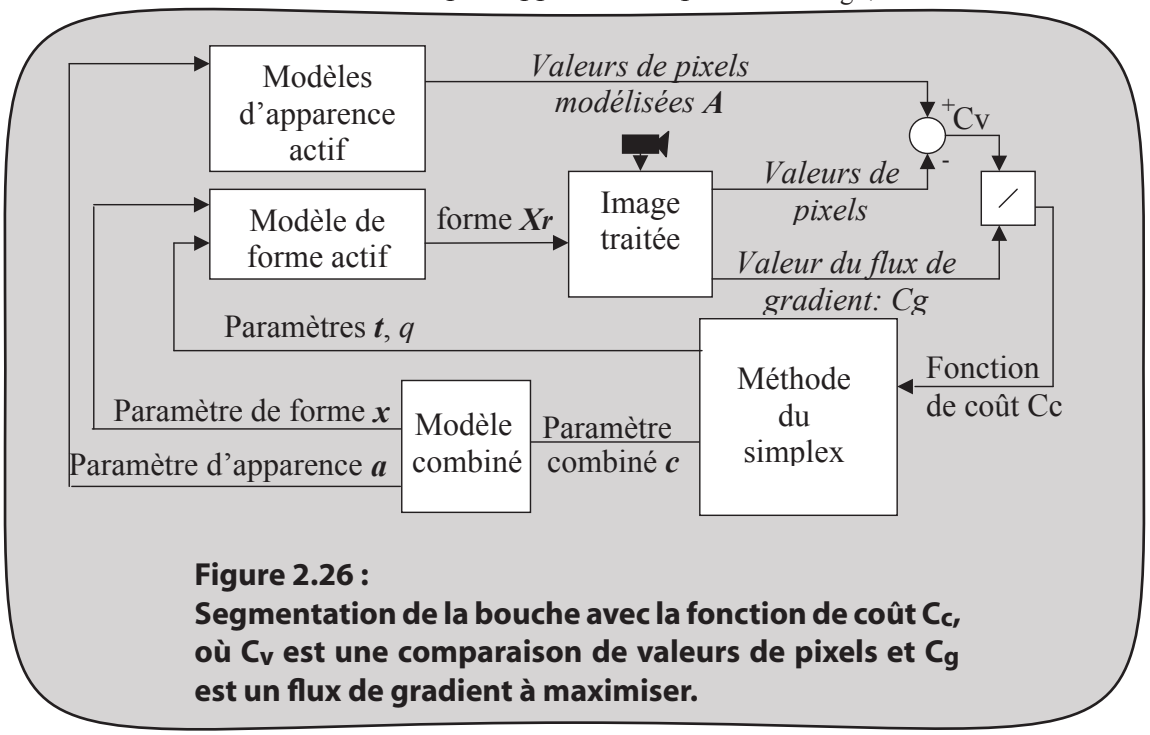


Figure 2.26 :
Segmentation de la bouche avec la fonction de coût C_c, où C_V est une comparaison de valeurs de pixels et C_g est un flux de gradient à maximiser.

a par exemple moins de chance d'être confondu avec le contour lèvre/dent grâce aux informations colorimétriques des pixels).

La figure 2.26 présente le principe de cette fonction de coût dans le cas mono-locuteur.

2.4.3.3 Fonction de coût utilisant des descripteurs locaux

Motivation

Si les fonctions C_g et C_v sont classiques dans la littérature et si la deuxième peut donner de bons résultats, nous avons voulu développer une méthode qui reposerait sur une description de l'apparence moins sujette au bruit que l'utilisation directe des valeurs YCbCr des pixels et qui considérerait donc un voisinage proche du pixel d'intérêt. En outre, nous voulions que notre méthode soit capable de gérer le comportement non-linéaire de l'intérieur de la bouche avec davantage de précision qu'une approche par modèle actif basée sur une ACP et donc intrinsèquement linéaire.

La solution retenue est d'utiliser une fonction de coût basée sur la comparaison des réponses de descripteurs locaux d'apparence à la prédiction de cette réponse, approche qui combine, en outre, à la fois des critères de forme et d'apparence.

Principe

Le principe est de prédire pour un jeu de paramètres de forme donné la réponse des descripteurs locaux et de la comparer avec la réponse réelle des descripteurs calculée sur l'image. La prédiction sera effectuée par un réseau de neurones de type perceptron multi-couches bien adaptés à la description du comportement de la zone intérieure de la bouche qui présente de nombreuses non-linéarités (bouche ouverte ou fermée, apparition ou disparition des dents). Les critères de type C_g, souvent retenus pour les opérations de segmentation, échouent généralement à gérer l'intérieur de la bouche car les non-linéarités annulent ou modifient la direction des vecteurs gradients selon les configurations.

La réponse des descripteurs dépendra bien entendu du mouvement des lèvres, entre autre dû à la parole (variabilité intrinsèque ou intra-locuteur), de l'identité du locuteur (variabilité extrinsèque ou inter-locuteur) ainsi que des conditions d'éclairage.

Nous avons choisi d'utiliser des filtres dérivés gaussiens comme descripteurs locaux (tels que présentés en 2.2).

Ces filtres gaussiens ont également été utilisés par Odisio ([Odisio, 2005]) et présentent une solution moins coûteuse à mettre en oeuvre qu'une utilisation d'un banc de filtres de Gabor telle que dans [Zhang, 2003] où chaque point de contrôle est décrit par 40 coefficients correspondant aux réponses des filtres, là où notre méthode donne des résultats satisfaisants en utilisant seulement trois coefficients par plan de l'espace couleur.

La taille des fenêtres de convolution sera d'un dixième de la largeur de bouche et elles seront centrées sur les points de contrôle de la forme labiale. Ainsi les descripteurs sont régulièrement répartis sur les contours et leurs recouvrements respectifs sont limités autant que possible.

Pré-filtrage rétinien

Comme nous voulons prédire les réponses des filtres gaussiens à partir de la forme, celles-ci devraient donc idéalement ne dépendre que de la forme. Cela est presque vrai dans le cas mono-locuteur qui nous intéresse puisque la variabilité inter-locuteur n'intervient alors pas encore. Néanmoins les changements d'éclairage peuvent modifier la réponse des filtres pour une même personne, même en l'absence de mouvement des lèvres.

Pour résoudre ce problème, nous supposons que les composantes CbCr sont suffisamment insensibles à ces variations et traitons Y avec le filtre rétinien (cf 2.1.3). Il est d'ailleurs à noter que du fait que la sortie du filtre rétine est une image de type gradient (les zones homogènes se retrouvent à des valeurs quasi nulles, les contours sont rehaussés), on ne pourrait pas obtenir de bons résultats en n'utilisant que les réponses des filtres sur cette luminance filtrée.

Une autre solution, presque équivalente, serait de donner en entrée, non pas les paramètres de formes, mais les paramètres du modèle combiné. Les variations d'éclairage seraient gérées par le modèle d'apparence qui adapterait la prédiction des filtres en conséquence. Néanmoins cela serait au prix d'un nombre supérieur de valeurs à optimiser. En pratique, les deux approches étant équivalentes en terme de précision, la moins consommatrice d'itérations a été jugée préférable.

Associateur forme/descripteurs locaux

Pour accomplir la meilleure prédiction possible des 9 réponses des filtres (3 filtres et 3 composantes) à partir de la forme, l'associateur doit être capable de tenir compte des non-linéarités du problème (car les réponses des filtres évoluent non-linéairement en fonction de l'ouverture de la bouche et des apparitions des dents, par exemple).

Nous avons donc choisi d'utiliser un réseau de neurones de type perceptron multi-couche, à une couche cachée, entraîné par rétropropagation à partir de la base d'apprentissage. L'Annexe 3 rappelle brièvement le principe des réseaux de neurones et de la rétropropagation.

Appliqué à notre méthode, le réseau de neurones va avoir une fonction d'approximateur universel et va donc, pour un jeu de paramètres du modèle de forme donné, prédire la réponse des descripteurs locaux d'apparence sur l'image traitée.

Pour entraîner le réseau, nous avons à notre disposition notre base d'images d'apprentissage. La forme (c'est-à-dire les coordonnées des points de contrôle manuellement annotés) étant connue pour toutes les images, on peut calculer la réponse des filtres gaussiens lorsque ceux-ci sont correctement placés vis à vis des contours, ce qui constitue notre vérité terrain.

En outre, on peut facilement obtenir les paramètres du modèle de forme correspondant à chaque image, en inversant l'équation du modèle de forme :

$$\mathbf{x}_i = \mathbf{P}_\mathbf{x}^T \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}} \right), \; 1 \le i \le N \qquad \text{(eq. 2.30)}$$

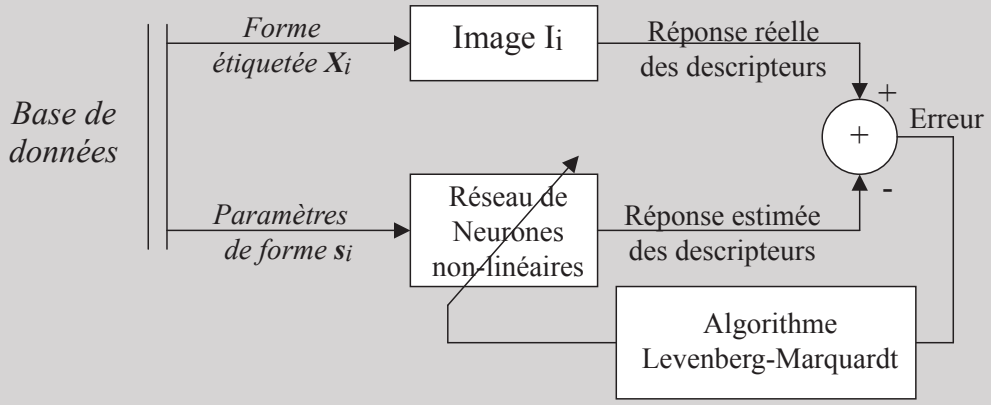


Figure 2.27 :
Principe de l'apprentissage de notre réseau de neurones.

Le principe de l'apprentissage (qui met en oeuvre l'algorithme d'optimisation de Levenberg-Marquardt, [Hagan, 1994]) est résumé par la figure 2.27.

Après avoir testé différents types de réseau, nous avons décidé de nous fixer sur un réseau à deux couches qui donne de bonnes performances pour une complexité raisonnable. Les fonctions de transfert de la première couche sont des tangentes hyperboliques tandis que celles de la deuxième couche sont des fonctions linéaires (afin que les sorties puissent prendre n'importe quelles valeurs).

Le réseau aura 6 entrées correspondant aux paramètres du modèle de forme.

Les sorties sont les réponses des descripteurs locaux mais il s'agit de vecteurs de grandes dimensions : nous avons en effet 20 points d'intérêt, 3 filtres et 3 composantes, soit 180 valeurs. Par souci de diminuer la taille du réseau de neurones, une ACP est effectuée sur l'espace des réponses des filtres afin de réduire la dimension du problème et le nombre de sorties de l'associateur. En conservant 95% de la variance, nous parvenons à réduire le nombre de sorties à 15 ce qui sera également le nombre d'unités cachées de la couche intermédiaire (voir figure 2.28 pour la structure de notre associateur non-linéaire).

Du fait que l'on doit respecter un rapport d'au minimum 5 à 10 entre le nombre de données et de paramètres et considérant que les approximateurs universels donnent de meilleurs résultats avec une sortie unique, nous avons finalement opté pour l'utilisation de 15 réseaux de petites tailles. Nous avons ensuite procédé par essai-erreur pour déterminer le nombre de coefficients de chaque couche et finalement opter pour un nombre intuitif d'unités cachées égal au nombre d'entrées.

Chacun des 15 réseaux a donc finalement 6 entrées, 6 unités cachées sur la couche intermédiaire et une sortie. La couche intermédiaire correspond donc à 6x6=36 paramètres. Comme la couche de sortie correspond à 6 valeurs, on doit donc déterminer 42 paramètres pour chaque réseau. Les données étant au nombre de N=400, le risque de surapprentissage est donc écarté.

Il est à noter que du fait de l'ACP sur les réponses des descripteurs locaux, celles-ci sont donc modélisées de manière linéaire de même que la forme. En revanche, la relation établie

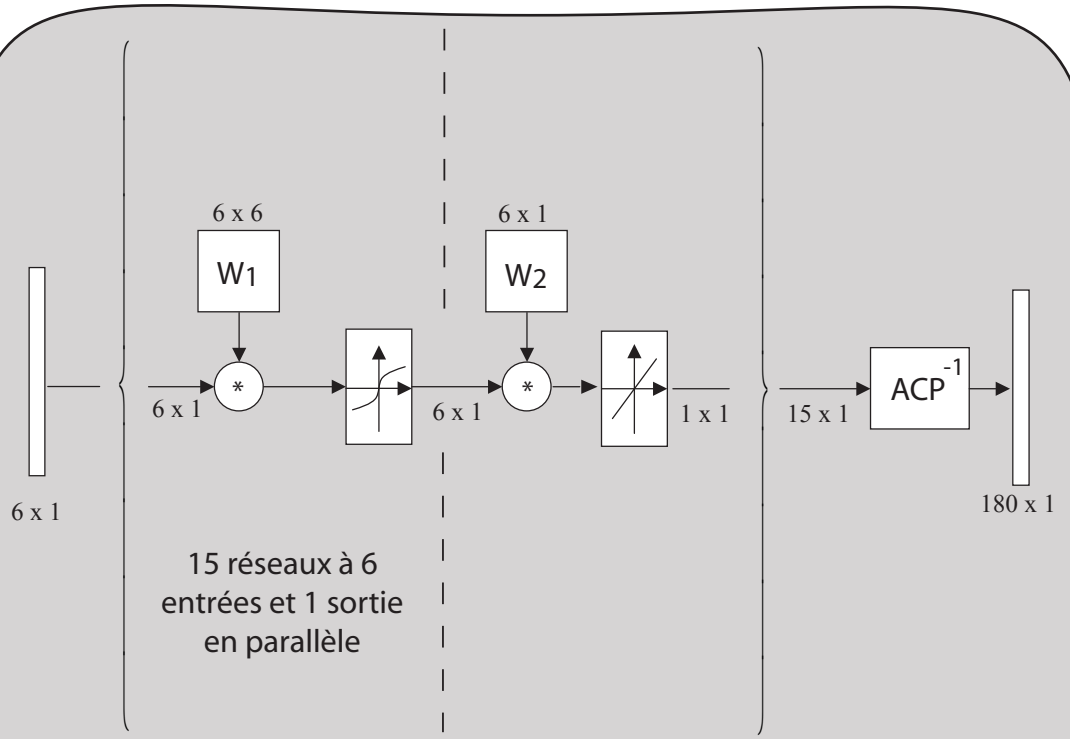


Figure 2.28 :
Structure de notre associateur non-linéaire.
Il s'agit d'une série de 15 réseaux de neurones non-linéaires à deux couches prédisant l'apparence locale des 20 points de lèvres à partir des paramètres de formes. Les entrées des réseaux sont d'abord multipliées par la matrice W_1 des poids de la première couche puis chaque poids obtenu est traité par une fonction tangente-hyperbolique. Ensuite, la sortie de la première couche est multipliée par la matrice de poids de la seconde couche qui est alors traitée par une fonction linéaire. On obtient ainsi les 15 sorties des réseaux qui sont projetées dans l'espace des réponses des descripteurs locaux par ACP inverse.

par notre associateur entre la forme et la description de l'apparence est non-linéaire.

Si l'on effectue la prédiction des réponses des filtres sur la base d'apprentissage à partir des $\boldsymbol{x}_i$, on observe que 87% de la variance réelle des filtres est restituée. La figure 2.29 illustre le gain de l'association non-linéaire : si l'on voulait trouver les réponses des filtres par régression linéaire à partir des $\boldsymbol{x}_i$ on arriverait seulement à 76% de la variance restituée. Si on rajoutait ensuite les valeurs des paramètres du modèle d'apparence, il faudrait 14 paramètres pour atteindre le même niveau de variance expliquée.

Dans un cas mono-locuteur, pour un jeu de paramètres de forme $\boldsymbol{x}$ donné, la fonction de coût C_f est définie comme la différence quadratique pondérée entre les réponses prédites des filtres connaissant $\boldsymbol{x}$ et les réponses observées sur l'image traitée aux emplacements définis par $\boldsymbol{x}$. La figure 2.30 résume le principe d'une segmentation mettant en oeuvre un tel critère.

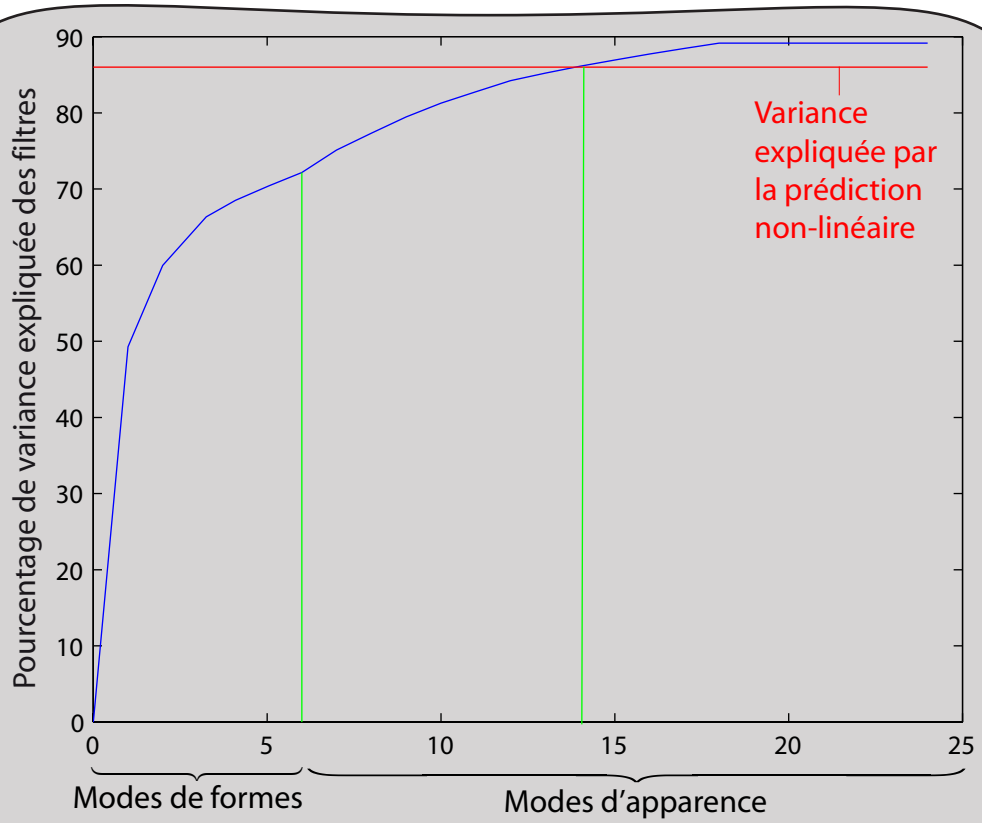


Figure 2.29 :
Evolution du pourcentage de variance expliquée des filtres par une régression linéaire faite à partir des modes de forme puis d'apparence.
En rouge : niveau de variance expliquée par la prédiction non-linéaire à partir des paramètres de forme.

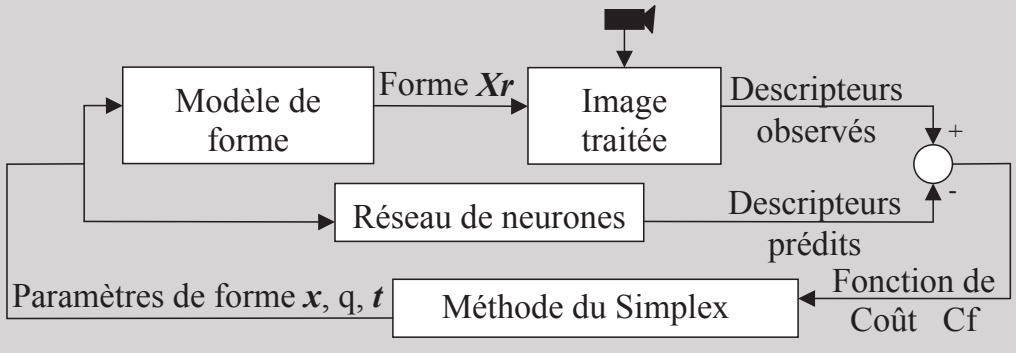


Figure 2.30 :
Principe de segmentation mono-locuteur avec descripteurs locaux.

2.4.4 Comparaison des fonctions de coût dans un cas mono-locuteur

Nous avons introduit quatre fonctions de coût dont nous allons à présent comparer les performances :

- C_g : fonction basée sur la maximisation du flux du gradient à travers les contours de la forme. Elle optimise le modèle de forme et donc les 6 valeurs du vecteur de paramètre $\boldsymbol{x}$ (cf figure 2.25).
- C_v : fonction basée sur la comparaison entre les pixels modélisés par le modèle d'apparence et les pixels observés sur l'image traitée. Elle optimise le modèle combiné forme/apparence et donc les 10 valeurs du vecteur de paramètre $\boldsymbol{c}$ (cf figure 2.26).
- C_c : fonction combinant les critères C_g et C_v. Elle optimise le modèle combiné forme/apparence et donc les 10 valeurs du vecteur de paramètre $\boldsymbol{c}$ (cf figure 2.26).
- C_f : fonction basée sur la comparaison entre des descripteurs locaux d'apparence et leur prédiction effectuée par un réseau de neurones en fonction de la forme. Elle optimise le modèle de forme et donc les 6 valeurs du vecteur de paramètre $\boldsymbol{x}$ (cf figure 2.30).

Le tableau 2.1 rassemble les résultats de segmentation pour chacun de ces quatre critères. Comme nous voulons uniquement nous intéresser à l'évaluation par les différentes fonctions de coût des paramètres optimaux des modèles statistiques, la position de la bouche $\boldsymbol{t}$ et son échelle q sont connues au départ. La grande précision de la localisation des commissures dans un cas mono-locuteur conduirait de toute façon à des résultats presque identiques.

Dans le cas de C_f et C_g, le modèle utilisé étant celui de forme, six paramètres seront donc optimisés. Dans le cas de C_v et C_c, il s'agit du modèle combiné forme/apparence, et donc de dix paramètres.

Les résultats sont des erreurs moyennes de localisation des points de contrôle de la forme sur 100 images (non présentes dans la base d'apprentissage). Les images sont traitées indépendamment les unes des autres, sans suivi d'aucune sorte. Enfin l'algorithme du Simplex est à chaque fois initialisé sur le vecteur nul, les intervalles de recherche de chaque paramètre correspondant alors à trois fois l'écart-type.

Cas de C_g

En considérant les résultats, sans grande surprise, la fonction de coût C_g qui n'inclut pas de modélisation de l'apparence donne de très loin les moins bons résultats.

La méthode C_g échoue particulièrement à donner des résultats réellement robustes et précis pour la zone de l'intérieur de la bouche. Par rapport aux autres fonctions, la détection du contour extérieur de la bouche est également moins précise, probablement en partie à cause de la mauvaise détection de l'intérieur.

Lorsque la bouche est fermée, le contour intérieur de la bouche est difficile à détecter par une maximisation de flux de gradient puisqu'il sépare alors deux zones de caractéristiques YCbCr similaires. En outre, lorsque la bouche s'ouvre, la direction des vecteurs gradients est susceptible de changer et les frontières dents/lèvres peuvent, dans certains cas, être confondues avec les frontières dents/fond de la bouche.

Fonction de coût	contour extérieur	contour intérieur	contours des dents	tous points	nombre moyen d'itérations
C_g	2.3 ± 1.3	6 ± 3.5	6.2 ± 3.5	4.4 ± 2.5	19.2
C_v	1.4 ± 0.9	2 ± 1.1	2.2 ± 1.2	1.8 ± 1	31.1
C_c	1.4 ± 0.8	1.9 ± 1	2.1 ± 1.1	1.6 ± 1	22.5
C_f	1.4 ± 0.8	1.8 ± 0.9	1.8 ± 0.9	1.5 ± 0.8	9.8

Tableau 2.1 :
Comparaison des fonctions de coût dans un cas mono-locuteur.
C_f : prédiction de descripteurs non-linéaires, C_v : comparaison de pixels, C_g : maximisation de flux de gradient, C_c : combinaison de C_g et C_v. Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type.

En comparaison, l'utilisation des descripteurs locaux agit comme une maximisation de gradient "intelligente" et qui s'adapterait à la configuration de la bouche. Les réponses attendues des filtres gradients Gx et Gy (figure 2.14) changent selon l'état de la bouche et la présence ou l'absence de dents, le filtre moyen G permet de discriminer les différentes frontières entre elles et donc, de ne pas confondre les contours.

De même C_v et C_c, en modélisant les valeurs des pixels pour chaque zone (lèvre, peau, dent, fond de la bouche) peuvent s'adapter aux problèmes de la zone intérieure avec bien plus de précision.

Comparaison de C_c et C_v

La fonction de coût C_c étant la combinaison des critères C_g et C_v, on constate qu'elle est plus précise que ces critères pris séparément. Par rapport à C_v, l'apport de l'information des champs de gradients permet à la fonction C_c de faire «coller» la forme aux contours en plaçant les points de contrôle sur des zones de forts gradients. Si le gain en précision est relativement limité, bien que réel, le gain en nombre d'itérations est, par contre, particulièrement conséquent (une dizaine), ce qui justifierait le temps consacré au calcul des champs de gradients.

Comparaison de C_f et C_c

Ces deux fonctions de coût comprennent une modélisation de l'apparence ainsi qu'un prise en compte des dérivées de l'image (par l'intermédiaire des champs de gradient pour C_c et par celui des filtres dérivés gaussiens du premier ordre pour C_f). En revanche, l'association entre forme et apparence est linéaire dans un cas (C_c) et non-linéaire dans l'autre (C_f).

Nous pouvons voir, d'après les résultats, que la fonction de coût C_f donne les meilleurs résultats et converge vers la solution plus rapidement que C_c.

Alors que le contour extérieur est détecté de manière quasiment équivalente dans les deux cas de figure, la prédiction non-linéaire des descripteurs pour C_f semble gérer avec davantage de précision la zone de l'intérieur de la bouche que le critère C_c qui n'exploite qu'une description linéaire de l'apparence dans cette zone.

En outre, comme dans le cas de C_f seuls les paramètres de forme doivent être optimisés, la convergence se révèle également bien plus rapide. De plus, l'information donnée par les filtres gaussiens est également beaucoup moins sensible au bruit que celle donnée directement par les pixels et les champs de gradient dans le cas de C_c : il est donc plus rapide d'atteindre un minimum et cela tend également à réduire le nombre d'itérations nécessaires.

En conclusion, la fonction de coût C_f donne les meilleurs résultats dans le cas d'une tâche mono-locuteur.

2.4.5 Discussion sur l'apport des descripteurs non-linéaires

Nous avons ensuite validé la pertinence de cette solution C_f en comparant ses performances de segmentation à celles d'autres méthodes, en étudiant les améliorations apportées respectivement par l'utilisation des filtres gaussiens pour représenter l'apparence dans la fonction de coût (par rapport à une description de l'apparence par les pixels) et par l'utilisation d'un réseau de neurones non-linéaires pour faire le lien entre la forme et l'apparence (par rapport à une association linéaire comme dans les AAMs classiques).

Dans toutes les méthodes testées, la forme est décrite par un même modèle actif obtenu par une ACP. L'apparence, quant à elle, est décrite de deux manières différentes:

- par des descripteurs locaux gaussiens (méthode A)
- par les valeurs YCbCr extraites sur une grille d'échantillonnage déduite de la forme (méthode B)

Pour chacun de ces deux cas, nous avons testé trois associateurs différents :1) un modèle d'apparence actif classique : ACP sur la description de l'apparence puis ACP pour combiner les modèles de forme et d'apparence de type (Cootes, Edwards et al. 2001), 2) l'apparence prédite à partir de la forme par une régression linéaire (comme dans [Odisio, 2005]), 3) l'apparence prédite à partir de la forme par un réseau de neurones non-linéaire.

Il est à noter que pour les associateurs 2) et 3) le nombre de paramètres à optimiser est inférieur à l'associateur 1) puisque seuls les 6 paramètres du modèle de forme doivent être optimisés (contre 10 pour le modèle combiné).

Les résultats de segmentation sur 100 images (absentes de la base d'apprentissage), avec position des commissures connues (afin de ne comparer que l'évaluation des paramètres des modèles par les fonctions de coût), sont donnés dans le Tableau 2.2. On notera enfin que la méthode A)3) correspond à notre fonction de coût Cf tandis que la méthode B)1) correspond à la fonction C_v.

Les résultats montrent que les meilleurs résultats sont bel et bien obtenus avec une paramétrisation par descripteurs locaux et un associateur non-linéaire.

Pour les méthodes de type B) (apparence décrite par les valeurs des pixels) les résultats sont similaires pour les trois associateurs. L'utilisation de la grille d'échantillonnage réduit

méthode	contour extérieur	contour intérieur	contours des dents	tous points	nombre moyen d'itérations
A) 1)	1.6 ± 1.1	2.4 ± 1.7	2.6 ± 1.8	2.1 ± 1.5	29.8
2)	1.7 ± 1.2	2.9 ± 1.8	3.1 ± 2.2	2.4 ± 1.7	17.2
3) (coût C_f)	1.4 ± 0.8	1.8 ± 0.9	1.8 ± 0.9	1.5 ± 0.8	9.8
B) 1) (coût C_v)	1.4 ± 0.9	2 ± 1.1	2.2 ± 1.2	1.8 ± 1	31.1
2)	1.5 ± 1	2.2 ± 1.3	2.5 ± 1.4	2 ± 1.2	15.1
3)	1.5 ± 0.9	2.1 ± 1.3	2.4 ± 1.3	2 ± 1.1	15.5

Tableau 2.2 :
Comparaison de méthode pour l'apport des descripteurs locaux et de l'associateur non-linéaire.
Méthodes: A) apparence décrite par les filtres gaussiens, B) apparence décrite par les valeurs des pixels,
1) modèle d'apparence actif, 2) apparence prédite à partir des paramètres de forme par régression linéaire, 3) apparence prédite à partir des paramètres de forme par un réseau de neurones non-linéaires.
Les résultats sont des erreurs de localisation et sont donnés en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type

en fait les problèmes de non-linéarités grâce à sa description précise de l'intérieur de la bouche qui est la zone où ce genre de problème est susceptible de se produire. La méthode B)1) donne les meilleurs résultats mais requiert un nombre supérieur d'itérations en raison de la quantité plus importante de paramètres à optimiser.

Les méthodes de type A) utilisant les descripteurs locaux gaussiens ne donnent de résultats réellement intéressants que pour la méthodologie 3) où un associateur non-linéaire est utilisé pour faire le lien entre la forme et l'apparence. En effet, pour les méthodes A)1 et A)2) la relation linéaire entre la forme et l'apparence entraîne une description pauvre de l'intérieur de la bouche alors que la prédiction non-linéaire des descripteurs prend en compte ces problèmes avec efficacité.

Ainsi l'utilisation d'un associateur non-linéaire et de descripteurs gaussiens ne semble pertinente que si ces deux options sont utilisées conjointement.

Enfin, si l'on compare les méthodes C_f (A)3)) et C_v (B)1)), nous pouvons voir d'après les résultats que la fonction de coût C_f donne les meilleurs résultats et converge vers la solution plus rapidement que C_v, conformément à ce qui avait été observé en 2.4.5. Cela est entre autre dû au fait que les filtres dérivés Gx et Gy donnent une mesure comparable à un champ de gradient, assurant que les contours détectés collent au plus près du contour.

2.5. Méthode retenue pour le cas mono-locuteur

2.5.1 Implémentation pratique de la méthode de segmentation de la bouche

Nous présentons ici la simple application du principe d'optimisation présenté dans le cas général au 2.4.2.3 à la fonction de coût présentée au 2.4.3.3. Le schéma de principe de cette segmentation est résumé par la figure 2.31.

Nous voulons obtenir le meilleur jeu de paramètres afin que notre modèle opère la meilleure segmentation possible de la bouche.

Soit $Cf(I_n)(\boldsymbol{x},q,\boldsymbol{t})$ la valeur de la fonction de coût Cf appliquée à l'image I_n pour les vecteurs de paramètres de forme $\boldsymbol{x}$, pour l'échelle q et la position $\boldsymbol{t}$.

Afin d'améliorer la vitesse de convergence et de diminuer le risque d'obtenir un minimum local, nous allons définir la meilleure initialisation possible pour l'algorithme ainsi que des intervalles de recherche appropriés pour les paramètres.

Première image I_1 d'une séquence

Le modèle de commissure est utilisé afin de détecter ces deux points clés. Cela donne des valeurs initiales généralement fiables pour la position de la bouche $\boldsymbol{t}$ ainsi que pour son échelle q.

Afin d'obtenir une estimation initiale de $\boldsymbol{c}$, nous testons l'Etat Général de la Bouche (EGB). Pour cela, on calcule pour chaque EGB: $Cf(I_1)(\boldsymbol{x}_j^{egb},q,\boldsymbol{t})$ pour $1 \leq j \leq 4$, les $\boldsymbol{x}_j^{egb}$ étant les jeux de paramètres moyens calculés lors de l'apprentissage. L'indice du minimum donne l'initialisation du DSM : $(\boldsymbol{x}_{jm}^{egb},q,\boldsymbol{t})$.

Nous procédons ensuite à la minimisation de $Cf(I_1)(\boldsymbol{x}_{jm}^{egb},q,\boldsymbol{t})$ par DSM, les intervalles de recherches pour le vecteur de paramètres $\boldsymbol{x}$ étant définis par l'EGB et les limites de paramètres $\boldsymbol{x}_{jm}^{b}$ (les $\boldsymbol{x}_j^{b}$ ayant été calculées lors de l'apprentissage (2.3.2.3)). Les intervalles équivalent à un vingtième de la largeur de la zone visage pour q et $\boldsymbol{t}$.

Après convergence du DSM, le jeu de paramètres final fourni par la convergence est noté : $(\boldsymbol{x}_1,q_1,\boldsymbol{t}_1)$.

Suivi

Nous cherchons tout d'abord à savoir si la bouche a peu bougé par rapport à l'image précédente. Pour l'image I_{n+1}, nous vérifions donc si :

$$\left| \frac{C_f(\mathbf{I}_{n+1})(\boldsymbol{x}_n,q_n,\boldsymbol{t}_n) - C_f(\mathbf{I}_{n+1})(\boldsymbol{x}_n,q_n,\boldsymbol{t}_n)}{C_f(\mathbf{I}_{n+1})(\boldsymbol{x}_n,q_n,\boldsymbol{t}_n)} \right| \leq 20\% \qquad \text{(eq. 2.31)}$$

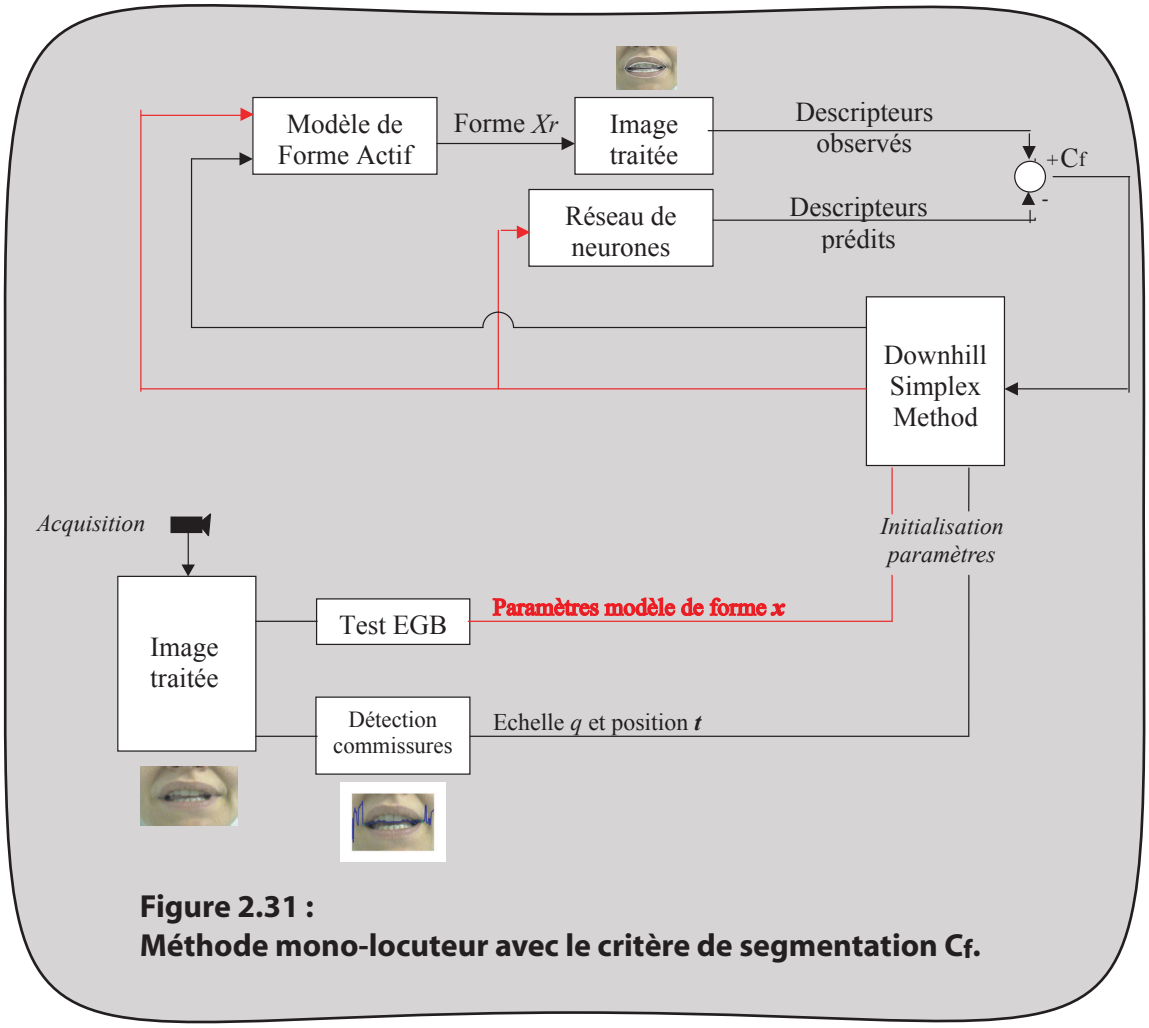


Figure 2.31 :
Méthode mono-locuteur avec le critère de segmentation C_f.

Si ce critère est vérifié, nous minimiserons donc $C_f(I_{n+1})(\boldsymbol{x},q,\boldsymbol{t})$ par DSM, avec $\boldsymbol{x}_n$ comme estimation initiale et des intervalles de recherche réduits correspondant à $0{,}5\sqrt{\lambda_k}$ pour le vecteur de paramètres $\boldsymbol{x}$ et q_n et $\boldsymbol{t}_n$ comme estimations initiales de q et $\boldsymbol{t}$.

Si le critère est non vérifié, nous testons à nouveau les EGB et détectons à nouveau les commissures, ce qui revient à repartir sur une image initiale.

2.5.2 Résultats quantitatifs de segmentation de la bouche :

Le tableau 2.3 donne la précision de la détection des contours sur environ 1800 images non présentes dans la base d'apprentissage (cinq séries complètes ainsi que leurs symétriques) avec détection automatique des commissures grâce au modèle local. Les scores légèrement meilleurs que ceux des Tableaux 2.1 et 2.2 pour les images initiales (c'est à dire sans suivi) proviennent de l'emploi des EGB pour initialiser le DSM (ce qui compense largement le fait que les commissures sont ici détectées et non connues). Enfin les scores en Suivi bénéficient de la connaissance de l'image précédente.

	contour extérieur	contour intérieur	contours des dents	tous points	nombre moyen d'itérations
Images indépendantes	1.2 ± 0.6	1.5 ± 0.7	1.5 ± 0.8	1.3 ± 0.7	7.6
Suivi	1.1 ± 0.5	1.4 ± 0.7	1.5 ± 0.8	1.2 ± 0.6	6.3

Tableau 2.3 :
Erreur de positionnement des points dans le cas mono-locuteur pour des images traitées indépendamment ou en Suivi.
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type

2.6 Bilan sur le cas mono-locuteur :

Nous avons présenté dans cette partie une application utilisant une fonction de coût non-linéaire adaptée à un cas mono-locuteur et donnant dans ce cadre des résultats à la fois précis, robustes et nécessitant peu d'itérations. La figure 2.32 donne quelques exemples de segmentation sur les images de la base de données.

La principale caractéristique de la fonction de coût retenue est qu'elle établit une association non-linéaire entre les descriptions de la forme et de l'apparence ce qui conduit à bien prendre en compte la zone intérieure de la bouche. En outre, l'utilisation des filtres gaussiens permet d'avoir une description de l'apparence prenant en compte les dérivées de l'image et peu sensible au bruit. Enfin, le recours à deux initialisations (détection des commissures et détermination de l'EGB) pour les paramètres à optimiser permet de rendre la convergence de l'algorithme de descente du simplex relativement rapide et robuste, en diminuant le risque de tomber sur des minima locaux de la fonction de coût.

Dans le cadre plus large de cette thèse et de son objectif multi-locuteurs, il faudra appliquer cette fonction de coût à des cas où le sujet peut varier et donc pour ceci tenir compte de la variabilité inter-locuteurs. En effet si le locuteur devient inconnu, la réponse des filtres n'est plus uniquement déterminée par la forme : d'une personne à l'autre, les couleurs de la peau et des lèvres, ainsi que le contraste entre les deux, sont susceptibles de changer.

Dans le chapitre suivant, nous introduirons les notions d'apparence statique (caractéristique d'un locuteur, qui correspondrait dans le cas mono-locuteur à l'apparence moyenne) et dynamique (fonction du mouvement et de l'éclairage et donc équivalente dans les faits à l'apparence utilisée dans le cas mono-locuteur).

Dans ce cadre élargi, les paramètres de l'apparence statique devront donc être également en entrée du réseau de neurones afin d'adapter la fonction de coût pour un usage multi-locuteurs. Lors de la segmentation d'une suite d'images du même locuteur, le fait de traiter séparément l'apparence statique permettra en outre de se ramener rapidement à un problème mono-locuteur dès que les paramètres de cette caractéristique auront convergé vers les valeurs adaptées au locuteur.

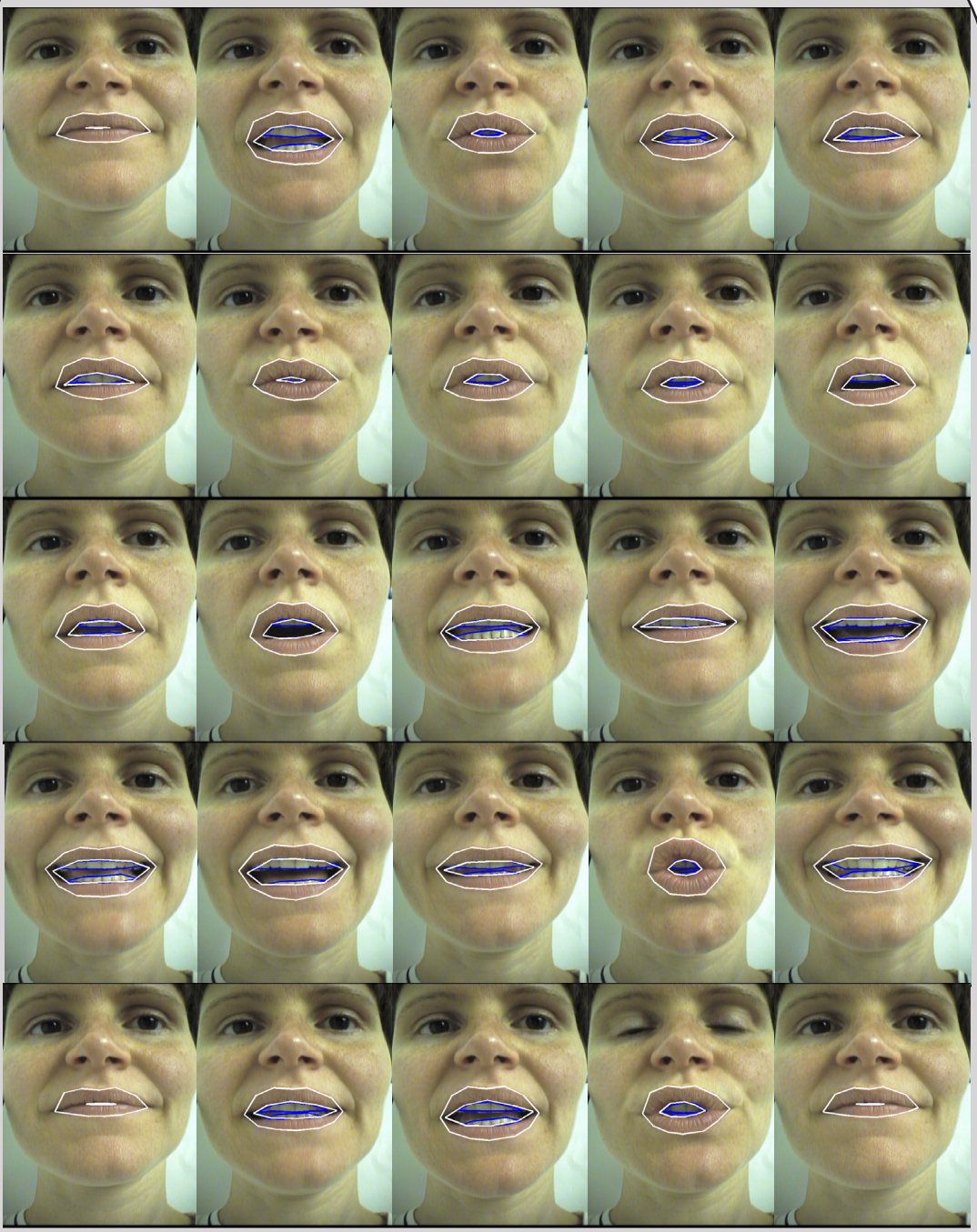


Figure 2.32 :
Exemples de segmentation de lèvres dans le cas mono-locuteur.

CHAPITRE 3

Modèle Multi-Locuteurs

Cette partie présente une généralisation des techniques présentées dans le chapitre 2 au cas multi-locuteurs. Le principal changement est qu'il faudra tenir compte de la variabilité d'apparence entre les personnes dont on veut segmenter les lèvres. Nous verrons que les conditions d'acquisition (principalement la caméra utilisée pour filmer la séquence vidéo traitée) doivent aussi être prises en compte pour avoir un modèle robuste.

3.1 Base de données

La base de données est constituée de vidéos de longueurs variables représentant douze locuteurs différents qui été ont enregistrées par trois caméras différentes dans des conditions d'éclairement diverses. Les deux premières caméras sont des mini-caméras fixées à 20 centimètres de la bouche avec des objectifs respectifs de 40 et 50 mm. Enfin,

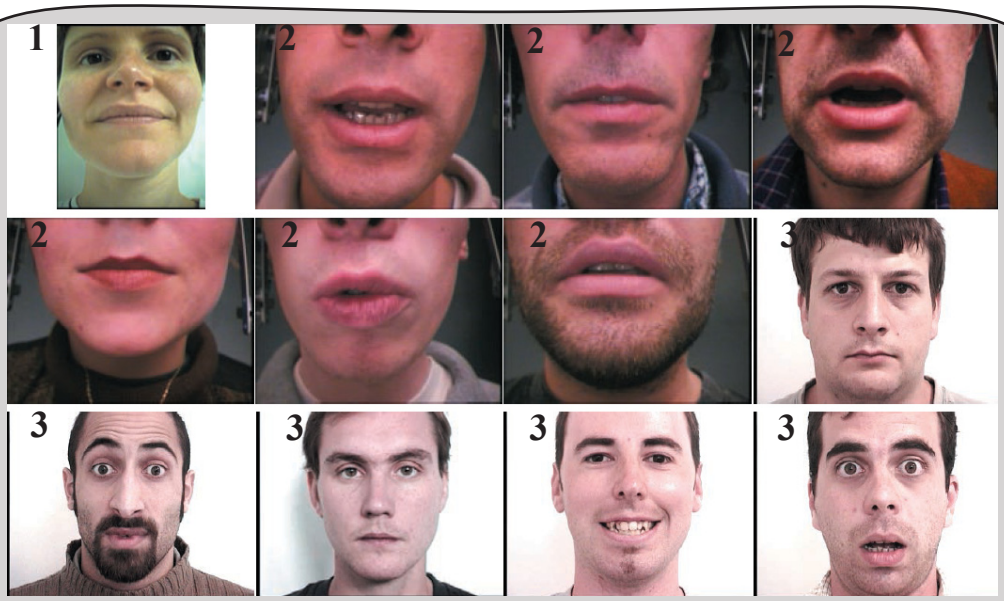


Figure 3.1 :
Exemples d'images de la base d'apprentissage pour chacun des douze locuteurs (avec la caméra correspondante indiquée à chaque fois).

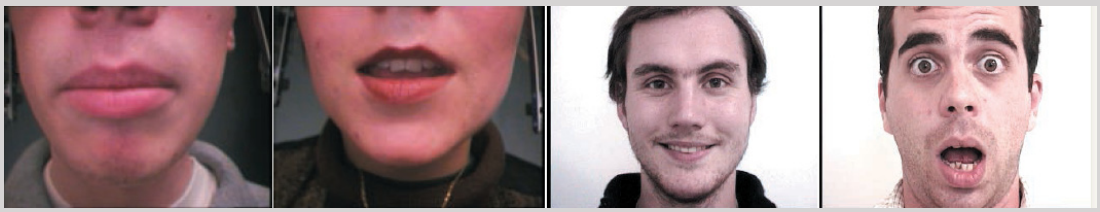


Figure 3.2 :
Exemples d'images typiques pour chaque EGB.
De gauche à droite: bouche fermée, ouverte, souriante, grande ouverte.

la troisième caméra est une webcam (voir figure 3.1).

450 images de cette base de donnée ont été sélectionnées pour faire partie de la base d'apprentissage de façon à offrir la plus large variété possible de formes de bouche. En outre, les symétriques verticales de ces 450 images sont également ajoutées si bien que nous aurons finalement N=900 images d'apprentissage.

Les images sont annotées de façon similaire au cas mono-locuteur avec 30 points de contrôle et l'opérateur assigne également un état général de la bouche (EGB) à chaque image. Enfin, l'opérateur choisit une image avec expression neutre pour chaque locuteur, cette image servant de référence pour l'échelle.

Dans notre travail, nous considérons que le visage est trouvé lors d'un pré-traitement et que l'on connaît donc la position de la bouche et de son voisinage (c'est à dire le bas du visage). Ainsi lors du test de notre algorithme, nous supposons que ce pré-traitement s'est déroulé et les images ont donc été recadrées quand cela était nécessaire autour du bas du visage. La figure 3.3 montre un exemple d'image de la base de donnée qui a été recadrée et segmentée.

3.2 Modèles

Cette partie présente les modèles statistiques utilisés par notre méthode de segmentation des lèvres. Le premier est un modèle colorimétrique de la lèvre et de la peau qui sera utilisé

Figure 3.3 :
Image de la base de donnée originale et image recadrée et segmentée.

principalement pour tenir compte de la variabilité du locuteur lors de l'initialisation des autres modèles. Un modèle de commissures et un modèle actif de la bouche sont également utilisés, ces derniers étant des extensions au cas multi-locuteurs de ceux présentés au chapitre 2.

3.2.1 Modèle colorimétrique de pixels

3.2.1.1 Principe

Ce modèle de couleur discrimine de façon grossière les pixels de peau et les pixels de lèvres. Il sera utilisé par la suite afin d'initialiser le modèle d'apparence de la bouche sur la première image d'une séquence (en adaptant les paramètres à la teinte du locuteur et à l'éclairage) et afin de rendre plus robuste la détection de commissure.

Sur un visage quelconque, on constate que les valeurs prises par les pixels de lèvres et de peau ont des distributions d'allures gaussiennes dans les espaces couleurs, ce qui incite naturellement à vouloir les décrire grâce à des mélanges de gaussiennes. Une approche similaire a déjà été suivie dans [Patterson, 2002] en utilisant la distribution des composantes RVB. Préférant utiliser l'information chromatique pour discriminer lèvre et peau, nous avons d'abord utilisé les composantes CbCr dans nos premiers travaux ([Gacon, 2005]), avant d'opter en définitive pour l'utilisation de l'image de pseudo-teinte **H**.

En effet, comme le montre la figure 3.4, les distributions statistiques de la peau et des lèvres ont davantage tendance à se recouvrir dans le plan CbCr que sur la pseudo-teinte, et ce, même dans un cas où les lèvres sont visuellement parfaitement séparables de la peau pour l'oeil humain.

En revanche si les valeurs de pseudo-teinte de la peau et des lèvres sont relativement stables pour une caméra, la figure 3.5 montre que l'histogramme de l'ensemble des images de l'apprentissage présente trois couples de gaussiennes correspondant à chacune des trois

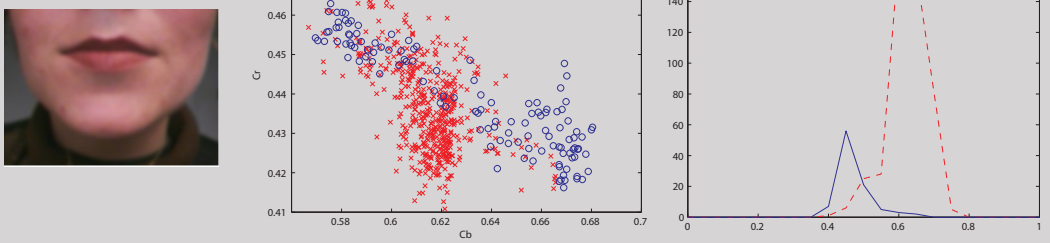


Figure 3.4 :
Comparaison des distributions des valeurs de peau et de lèvre sur le CbCr et la pseudo-teinte sur une image.
A gauche répartition des pixels dans le plan CbCr (peau/lèvre : x rouge/ o bleu).
A droite histogrammes en pseudo-teinte (peau/lèvre : pointillé rouge/ bleu).

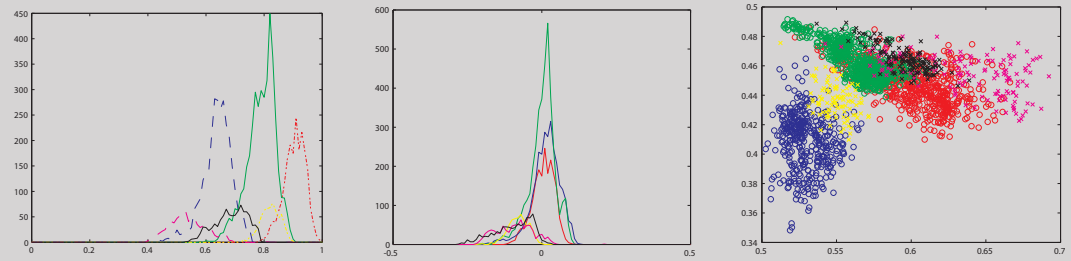


Figure 3.5 :
Distribution des valeurs colorimétriques sur l'ensemble de la base de donnée (900 images).
De gauche à droite : histogramme de la pseudo-teinte, histogramme de la pseudo-teinte recentrée sur le pic de la peau, répartition dans le plan CbCr.
Code Couleurs pour chaque caméra (peau/lèvre) :bleu/violet, vert/noir, rouge/jaune.

caméras présentes dans notre base de données (on observe un phénomène similaire en CbCr). Nous avons donc choisi de "recaler" les caméras en soustrayant la valeur du pic correspondant à la peau. La distribution de la pseudo-teinte est ainsi recentrée autour du pic de peau ce qui nous permet de construire un modèle colorimétrique qui sera moins sensible aux paramètres colorimétriques de la caméra.

Cela suppose évidemment de pouvoir déterminer avec précision la valeur du pic de pseudo-teinte de la peau du visage. La figure 3.6 montre que l'histogramme de la pseudo-teinte pour une image présente généralement au moins trois pics : un pour la peau, un petit pour les lèvres et un (ou plusieurs) autre(s) pic(s) pour le fond (plus ou moins important(s) selon la qualité du cadrage).

Sous l'hypothèse que l'image est centrée sur le bas du visage, supprimer les pics parasites peut s'accomplir facilement en réalisant un histogramme pondéré **hist_pond**, les pixels les plus éloignés du centre de l'image étant pénalisés selon la procédure suivante :

$$
\begin{aligned}
&\textit{pour } i = 1{:}\textit{largeur} \\
&\quad \textit{pour } j = 1{:}\textit{hauteur} \\
&\qquad \textit{val} = \text{valeur}\left(\text{H}\left(i, j\right)\right) \\
&\qquad \text{histo\_pond}\left(\textit{val}\right) = \text{histo\_pond}\left(\textit{val}\right) + \frac{1}{\left(j - \frac{\textit{largeur}}{2}\right)^2 + \left(i - \frac{\textit{hauteur}}{2}\right)^2} \\
&\quad \textit{fin} \\
&\textit{fin}
\end{aligned}
\qquad \text{(eq. 3.1)}
$$

où *largeur* et *hauteur* désignent la taille de l'image traitée, i et j les indices des lignes et colonnes et *val* la valeur de pseudo-teinte d'un pixel donné. La figure 3.6 montre le

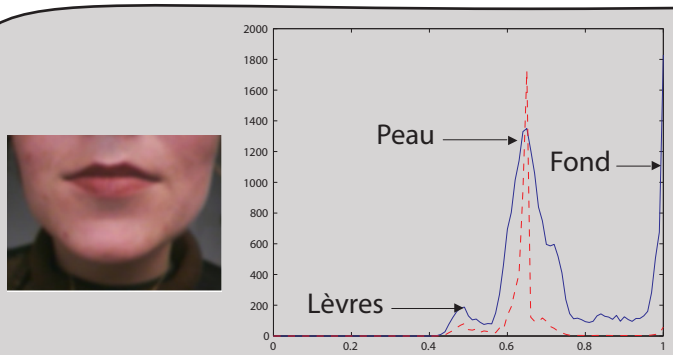


Figure 3.6 : Histogrammes de teinte du visage.
En pointillé rouge : histogramme pondéré afin de détecter le pic de la peau
En bleu: histogramme non pondéré.

résultat de cette procédure qui permet d'atténuer considérablement la contribution du fond de l'image.

La valeur du pic maximum de l'histogramme pondéré de pseudo-teinte est donc calculée puis soustraite aux valeurs prises par les pixels pour chacune des images de la base d'apprentissage. Les valeurs recentrées de **H** sont ensuite collectées sur chaque image étiquetée (en pratique, seul un dixième des valeurs disponibles ont été utilisées afin de réduire les temps de calcul) puis sont utilisées pour calculer la distribution d'une variable statistique.

Soit h la valeur de pseudo-teinte d'un pixel que l'on veut classer dans l'une des deux catégories possibles $\boldsymbol{c}_i$ avec $\boldsymbol{c}_1$: peau et $\boldsymbol{c}_2$: lèvre. La densité de probabilité d'appartenance de $\boldsymbol{h}$ à l'une des deux catégories est :

$$p\left(h \mid c_i\right)=\sum_{j=1}^{Kc_i} wc_{i,j} \times p\left(h \mid \left(\boldsymbol{\mu c}_{i,j}, \mathrm{S}\, \boldsymbol{c}_{i,j}\right)\right) \qquad \text{(eq. 3.2)}$$

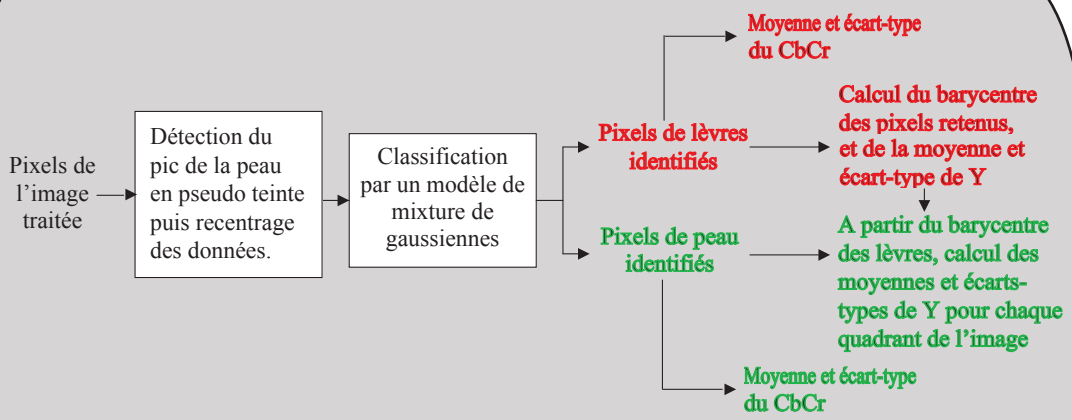
où Kc_i est le nombre de gaussiennes modélisant la catégorie et où $wc_{i,j}$, $\mu c_{i,j}$ et $\sigma_{i,j}$ (respectivement poids, moyenne et variance) ont les caractéristiques des gaussiennes avec $p(h,(\mu c_{i,j},o_{i,j}))$ telle que:

$$p\left(h \mid \left(\boldsymbol{\mu c}_{i,j}, \sigma_{i,j}\right)\right)=\frac{1}{\left(2\pi\sigma_{i,j}\right)^{1/2}} \exp\left(-\frac{1}{2}\left(h-\boldsymbol{\mu c}_{i,j}\right)^{T} \sigma_{i,j}\left(h-\boldsymbol{\mu c}_{i,j}\right)\right) \qquad \text{(eq. 3.3)}$$

la dimension du problème étant unique dans notre cas (la valeur de pseudo-teinte).

L'utilisation de l'algorithme d'optimisation EM et du principe de la MDL (cf. 2.3.1.2) conduit à $Kc_1=1$ et $Kc_2=1$ avec nos données.

Sans recentrages des données autour de la valeur du pic de la peau, on obtiendrait 3 gaussiennes pour chaque classe, correspondant à des distributions clairement distinctes des valeurs de pseudo-teinte pour chacune des caméras. De même, l'utilisation des composantes CbCr conduirait à avoir deux fois trois gaussiennes qui se recouvriraient grandement.



Après recentrage autour du pic de la peau, la distribution statistique des valeurs des pixels étant connue, la classification est obtenue en calculant pour chaque pixel s'il est plus susceptible d'appartenir à la classe 'peau' ou à la classe 'lèvres'. Les groupes de pixels de 'lèvres' éloignés du pic de minima de luminance sont ensuite supprimés dans un deuxième temps. Enfin, la connaissance du barycentre des points identifiés comme 'lèvres', permet de couper l'image en quadrant pour lesquels on calcule les valeurs moyennes des composantes YCbCr de chacun.

De gauche à droite: Image traitée, classification intermédiaire des pixels (blanc : lèvres, gris : peau, noir : non affecté), classification finale après retrait des blocs de lèvres "parasites".

De gauche à droite: Image traitée, classification des pixels (blanc : lèvres, gris : peau, noir : non affecté), image synthétique avec les valeurs YCbCr moyenne pour chaque quadrant.

Figure 3.7 :
Classification couleur des pixels.

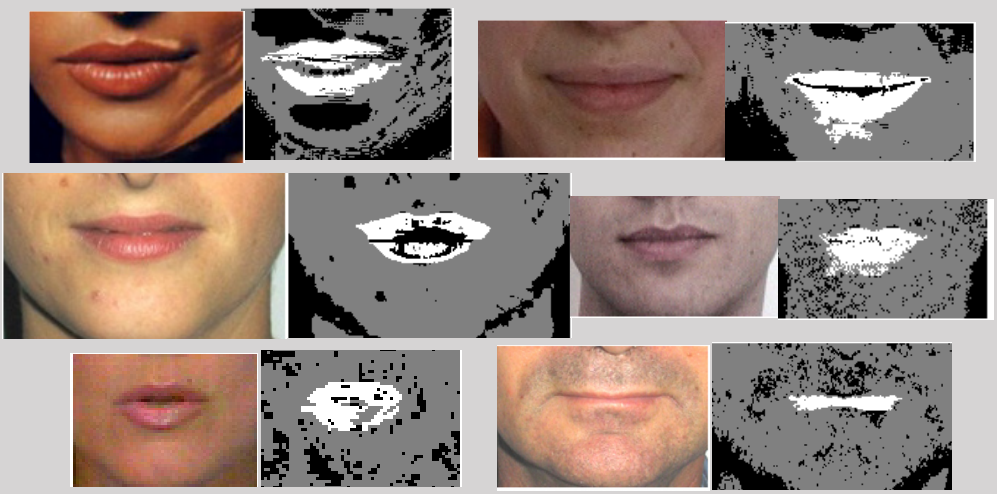


Figure 3.8 :
Exemple de classification de pixels pour des locuteurs et des conditions d'acquisition différents de la base d'apprentissage.
(blanc : lèvres, gris : peau, noir : non affecté)

3.2.1.2 Classification de pixels

La catégorie d'appartenance la plus probable pour le pixel est celle qui maximise $p(h|\boldsymbol{c}_i)$.

Afin de prendre en compte le fait que certains pixels n'appartiennent à aucune des deux catégories, nous avons fixé un seuil minimal acceptable pour la valeur de $p(h|\boldsymbol{c}_i)$. Cela assure que le pixel n'est pas trop éloigné de la distribution statistique. Le cas échéant il n'est assigné à aucune catégorie.

Ainsi, tous les pixels de l'image traitée sont soit identifiés comme étant de la peau ou des lèvres, soit ils restent sans affectation. Aucun critère de forme n'étant utilisé pour cette classification, il arrive que des blocs de pixels soient identifiés comme 'lèvres' alors qu'ils sont relativement éloignés de la zone de la bouche, ce cas de figure se présentant relativement souvent pour des pixels situés près du nez ou du cou. Ces pixels 'parasites' sont traités en supprimant les blocs de pixels trop éloignés du germe de minima de luminance (tel que définis au 2.3.1.1), le germe étant systématiquement détecté dans la zone de la bouche.

Comme nous voulons utiliser cette classification pour trouver la meilleure initialisation possible pour notre modèle d'apparence, nous calculons les valeurs moyennes et les écarts-types des composantes CbCr des pixels classifiés grâce au modèle de pseudo-teinte. Ainsi nous obtenons une approximation des caractéristiques colorimétriques des lèvres et de la peau d'un locuteur. En revanche, comme la luminance n'est généralement pas homogène sur le visage, les valeurs moyennes de la luminance sont calculées dans quatre quadrants de l'image (le découpage en quadrants étant déterminé par la position du barycentre des

lèvres) de façon à obtenir non seulement une information sur la chromacité mais également sur l'éclairage de la scène.

La méthode est entièrement résumée par la figure 3.7.

Ce modèle colorimétrique a été testé et implémenté dans une interface graphique ([Lopez Medina, 2006]) et détecte avec succès 70% des pixels de lèvres sur la base d'apprentissage avec environ 20% de fausses alarmes, performances suffisantes pour l'utilisation que nous souhaitons en faire. Il est enfin à noter qu'il donne des résultats satisfaisants même si le locuteur n'appartient pas à la base d'apprentissage (figure 3.8).

3.2.2 Modélisation locale des commissures des lèvres

Cette partie est grandement similaire à la méthode utilisée dans le cas mono-locuteur (2.3.1). L'objectif est de détecter les commissures des lèvres, points-clés permettant de déterminer la position et l'échelle de la bouche.

La ligne d'intérêt pour la recherche des commissures est toujours celle de minima de luminance construite de proche en proche à partir du germe initial détecté automatiquement (voir 2.3.1.1 et la figure 2.11).

Dans le chapitre 2.3.1, la détection s'effectue en calculant la réponse de filtres dérivés gaussiens sur les valeurs YCbCr de l'image. Néanmoins, la figure 3.5 a montré que les distributions de la peau et des lèvres étaient très différentes d'une caméra à l'autre et se recouvraient, ce qui conduirait à une diminution de la robustesse de la méthode si elle était appliquée sans aménagements au cas multi-locuteurs.

Ce problème peut toutefois être corrigé de façon similaire au modèle colorimétrique du 3.1 en recentrant les distributions des pixels autour des valeurs de la peau. Pour cela, on peut utiliser la classification fournie par le modèle colorimétrique qui permet en effet de détecter efficacement les pixels de peau. Or la figure 3.9 montre que les histogrammes des pixels de peau présentent pour les composantes CbCr deux pics très étroits tandis que la luminance a une distribution plus étendue.

Afin de recentrer les données sur celles de la peau, les valeurs du maximum des pics du CbCr et la moyenne de Y seront calculées puis soustraites aux composantes YCbCr pour chaque image de la base de données.

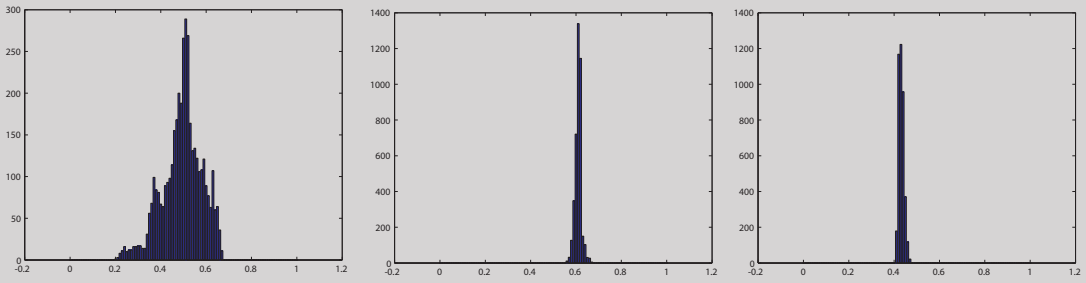


Figure 3.9 :
Histogrammes YCbCr des pixels de peau détectés par le modèle colorimétrique.

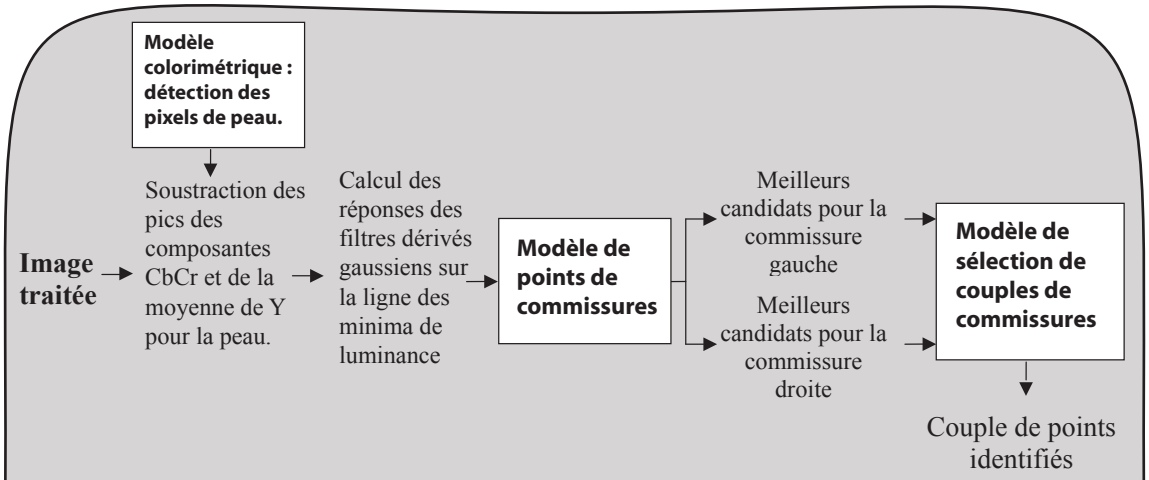


Figure 3.10 :
Détection des commissures des lèvres.
Après détection des pixels de peau par le modèle couleur et recentrage des valeurs YCbCr autour de la référence de la peau, les réponses des filtres dérivés gaussiens sont calculées sur la ligne des minima de luminance. A partir de ces réponses, un premier modèle statistique sélectionne les points les plus vraisemblables pour chaque commissure puis un second modèle désigne le couple de points le plus vraisemblable.

Les réponses des filtres gaussiens sont ensuite collectées sur les images recentrées et leur distribution statistique décrite par un mélange de gaussiennes construite selon le même principe qu'au 2.3.1. La figure 3.10 présente le principe de la méthode rendue invariante aux changements de caméras, le seul changement étant l'intervention du modèle colorimétrique pour calibrer les données YCbCr afin qu'elle soit traitée par le modèle de commissures.

3.2.3 Modèle actif de forme et d'apparence échantillonnée

3.2.3.1 Apparence statique et dynamique

On peut considérer que les variations d'apparence d'une image se divisent en des variations de nature inter-locuteur (c'est à dire liées à l'identité de la personne et à sa pigmentation), des variations de nature intra-locuteur (dues aux mouvements du visage, provoqués par la parole par exemple) et les variations d'apparence liées à l'environnement ou à la prise de vue (typiquement le changement d'éclairage ou de caméra).

Le modèle étant à présent multi-locuteurs, il faudra donc tenir compte des variations inter-locuteur qui avaient pu être ignorées dans le chapitre 2. L'objectif est donc de compléter la modélisation précédente en tenant compte de cette nouvelle dimension.

Nous avons pour cela envisagé de faire une distinction entre ce que nous avons appelé

l'apparence statique et l'apparence dynamique et de ne plus considérer uniquement l'apparence «globale» utilisée dans le cas mono-locuteur :

- l'apparence statique correspond à une description générique de l'apparence caractéristique du locuteur traité : typiquement la couleur des lèvres et de la peau (caractérisées par les composantes CbCr) mais également l'éclairage de la scène (information portée par la composante Y) et la caméra utilisée pour l'acquisition. Elle est estimée comme étant la moyenne de l'apparence d'un locuteur sur une série d'images.
- l'apparence dynamique est définie comme la différence entre l'apparence globale et l'apparence statique : elle correspond donc aux variations d'apparence dues au mouvement et à la parole.

Ainsi les variabilités d'apparence inter et intra-locuteur sont découplées. Par rapport au chapitre 2, l'apparence statique peut être vue comme un modèle de l'apparence moyenne d'un locuteur tandis que l'apparence dynamique est équivalente dans son usage à l'apparence globale du cas mono-locuteur.

Lors de la segmentation d'une suite d'images d'un même locuteur, le fait de traiter séparément l'apparence statique permettra de se ramener quasiment à un problème mono-locuteur dès que les paramètres de l'apparence statique auront convergé vers les valeurs adaptées au locuteur et à la scène.

Il est à noter que la variabilité de la forme pourrait également être considérée comme ayant une partie intra-locuteur et une partie inter-locuteur et ces deux variabilités pourraient donc être également découplées. Néanmoins, dans la pratique nous avons constaté que ce rajout de sophistication et de complexité (le nombre de paramètres à optimiser étant augmenté) dans la modélisation ne se traduisait pas en amélioration significative de la qualité de segmentation et cette approche a donc été abandonnée.

3.2.3.2 Apprentissage des données

Après avoir annoté la base d'apprentissage dans la partie 3.1, nous avons obtenu les vecteurs $\boldsymbol{X}_i$ ($1 \leq i \leq N$) qui contiennent les coordonnées des points définissant la forme. A partir de ces données, l'apparence échantillonnée est apprise pour chaque image en utilisant la grille d'échantillonnage selon la méthode définie en 2.3.2.1 et les valeurs des composantes YCbCr sont extraites en 728 pixels et sauvegardées dans les vecteurs de 2184 valeurs (728x3) notés $\boldsymbol{A}_i$ ($1 \leq i \leq N$).

L'apparence statique est alors calculée pour chaque série d'image d'un même locuteur (les séries originales et leurs symétriques étant considérées comme des séries indépendantes) comme étant la moyenne des $\boldsymbol{A}_i$ correspondants et est sauvegardée dans les vecteurs $\boldsymbol{S}_i$ ($1 \leq i \leq N$).

Pour chaque image on effectue ensuite la différence entre les $\boldsymbol{A}_i$ et les $\boldsymbol{S}_i$ ce qui fournit l'apparence dynamique sauvegardée dans les vecteurs $\boldsymbol{D}_i$ ($1 \leq i \leq N$).

Nous avons discuté au 3.2.1 et au 3.2.2 du fait que le changement de caméra et de conditions d'acquisition conduisait à des décalages des distributions des valeurs YCbCr. Dans le cas du modèle de commissures, cet inconvénient a été contourné en utilisant les informations fournies par la classification de pixels en classe peau/lèvre pour recentrer

les données sur les valeurs de la peau. Il est à noter que l'on pourrait suivre une approche similaire pour le modèle d'apparence statique en soustrayant la valeur de la moyenne de Y et des pics du CbCr à l'apparence $\boldsymbol{A}$ pour chaque image de la base d'apprentissage. Dans ce cas, l'apparence statique correspondrait principalement à la différence de teinte entre la peau et les lèvres et aux variations d'éclairement sur le visage (l'apparence dynamique restant quant à elle totalement inchangée). Nous verrons dans la suite que l'utilisation du modèle colorimétrique en prétraitement permet déjà au modèle de bouche de s'adapter efficacement aux trois caméras présentes dans l'apprentissage et en pratique les résultats se révèlent parfaitement équivalents sur les images de la base de données que l'on recentre les données ou non. Néanmoins, cela peut permettre de rendre le modèle applicable à des images prises dans des conditions différentes de l'apprentissage, comme nous le verrons au chapitre 4.

3.2.3.3 Analyse en composantes principales

Une fois les données collectées, nous procédons à des Analyses en Composante Principale avec les vecteurs $\boldsymbol{X}_i$, $\boldsymbol{D}_i$ et $\boldsymbol{S}_i$. Pour cela les valeurs moyennes ($\overline{\boldsymbol{X}}$, $\overline{\boldsymbol{D}}$ et $\overline{\boldsymbol{S}}$), les matrices de covariances ($\mathbf{Cov_x}$, $\mathbf{Cov_d}$ et $\mathbf{Cov_s}$) et leurs vecteurs propres ($\boldsymbol{p_{x,j}}$, $\boldsymbol{p_{d,n}}$ et $\boldsymbol{p_{s,n}}$) ainsi que les valeurs propres correspondantes ($\lambda_{x,j}$, $\lambda_{d,n}$, $\lambda_{s,n}$) sont calculés avec $1 \leq j \leq 60$ et $1 \leq n \leq 2184$.

L'étape suivante consiste en une réduction de dimension effectuée en ne sélectionnant que les vecteurs propres correspondant aux valeurs propres les plus élevées.

On garde 95% de la variance totale de la forme ce qui pour nos données correspond à 8 vecteurs propres qui sont placés dans la matrice $\mathbf{P}_X$. Pour l'apparence, nous conservons 90% de la variance dans les deux cas (statique et dynamique) ce qui correspond respectivement à 40 et 6 modes propres qui sont placés dans les matrices $\mathbf{P_d}$ et $\mathbf{P_s}$:

$$\begin{aligned} \boldsymbol{D} &= \overline{\boldsymbol{D}} + \mathbf{P}_\mathbf{d}\boldsymbol{d} \\ \boldsymbol{S} &= \overline{\boldsymbol{S}} + \mathbf{P}_\mathbf{s}\boldsymbol{s} \end{aligned} \qquad \text{(eq. 3.4)}$$

La figure 3.11 présente les 6 premiers modes de variations du modèle de forme. On constate que le premier mode correspond à un sourire et le deuxième à une protrusion et qu'ils sont relativement similaires aux modes du modèle mono-locuteur.

La figure 3.12 présente, quant à elle, l'effet des 6 modes de variations du modèle d'apparence statique (pour une forme et une apparence dynamique moyennes). Le premier mode contrôle le niveau d'éclairage tandis que le deuxième en contrôle l'orientation. Le troisième mode contrôle, quant à lui, le réglage colorimétrique de la caméra. Le quatrième correspond à la présence de barbe tandis que le cinquième et sixième correspondent à la pigmentation du locuteur et au contraste entre les lèvres et la peau.

Dans le cas mono-locuteur nous avions lié les variations de formes et d'apparence grâce à un modèle combiné. Dans le cas multi-locuteurs présent, comme la forme et l'apparence statique sont supposées indépendantes l'une de l'autre, c'est la forme et l'apparence

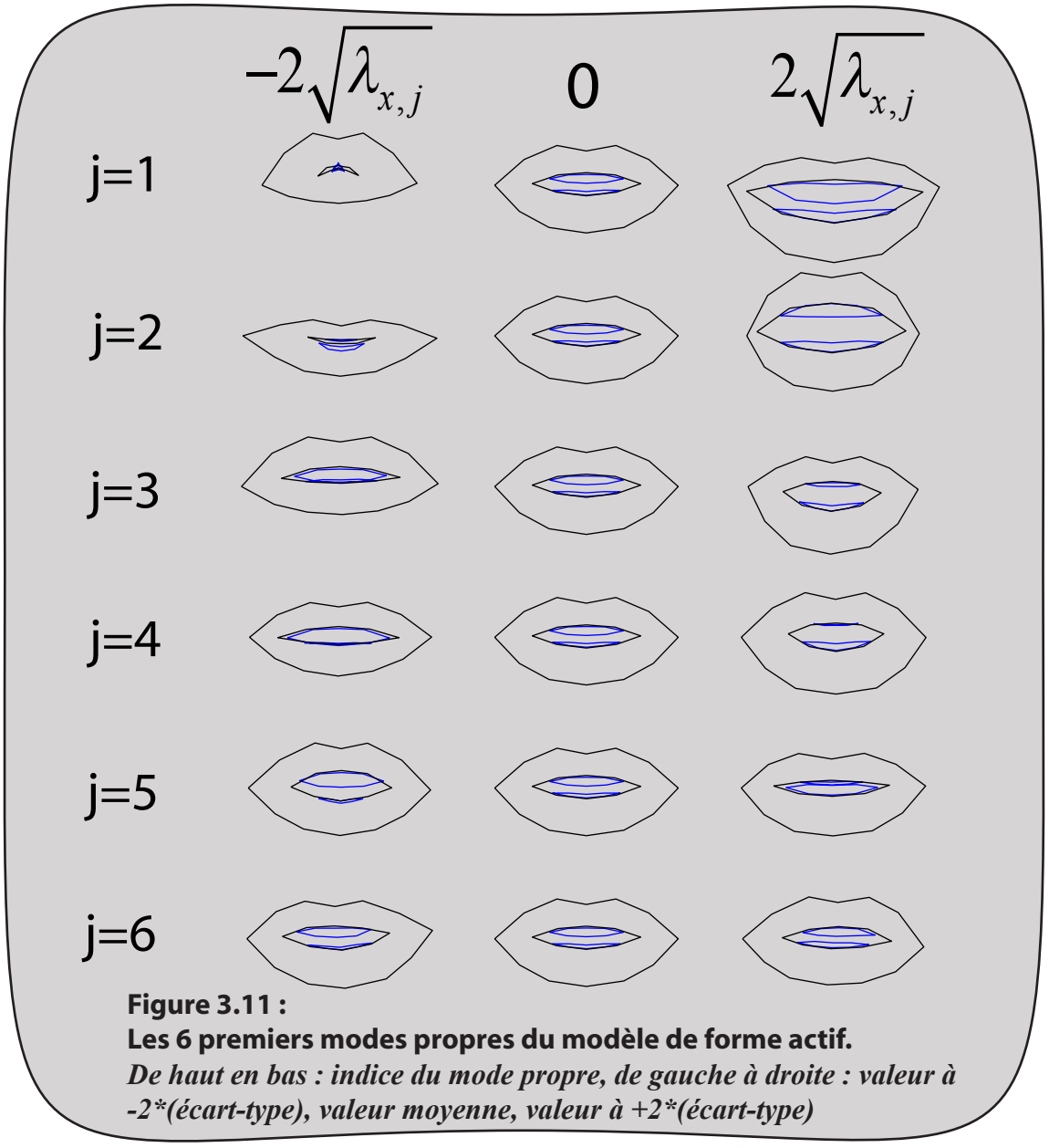


Figure 3.11 :
Les 6 premiers modes propres du modèle de forme actif.
De haut en bas : indice du mode propre, de gauche à droite : valeur à -2*(écart-type), valeur moyenne, valeur à +2*(écart-type)

dynamique qui seront modélisées conjointement.

Nous calculons donc pour chaque image les valeurs correspondantes des vecteurs de poids $\boldsymbol{x}_i$ (pour la forme) et $\boldsymbol{d}_i$ (pour l'apparence dynamique) avec $1 \leq i \leq N$.

$$\mathbf{x}_i = \mathbf{P}_\mathbf{x}^T \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}} \right) ; \mathbf{d}_i = \mathbf{P}_\mathbf{d}^T \left(\boldsymbol{D}_i - \overline{\boldsymbol{D}} \right) \qquad \text{(eq. 3.5)}$$

Nous procédons alors à une nouvelle ACP avec les valeurs de ces paramètres de poids afin d'obtenir un modèle statistique qui lie les variations de forme (représentées par les $\boldsymbol{x}_i$) et les variations d'apparence dynamique (représentées par les $\boldsymbol{d}_i$), ceci afin d'obtenir une modélisation conjointe et cohérente des deux grandeurs qui sont supposées être

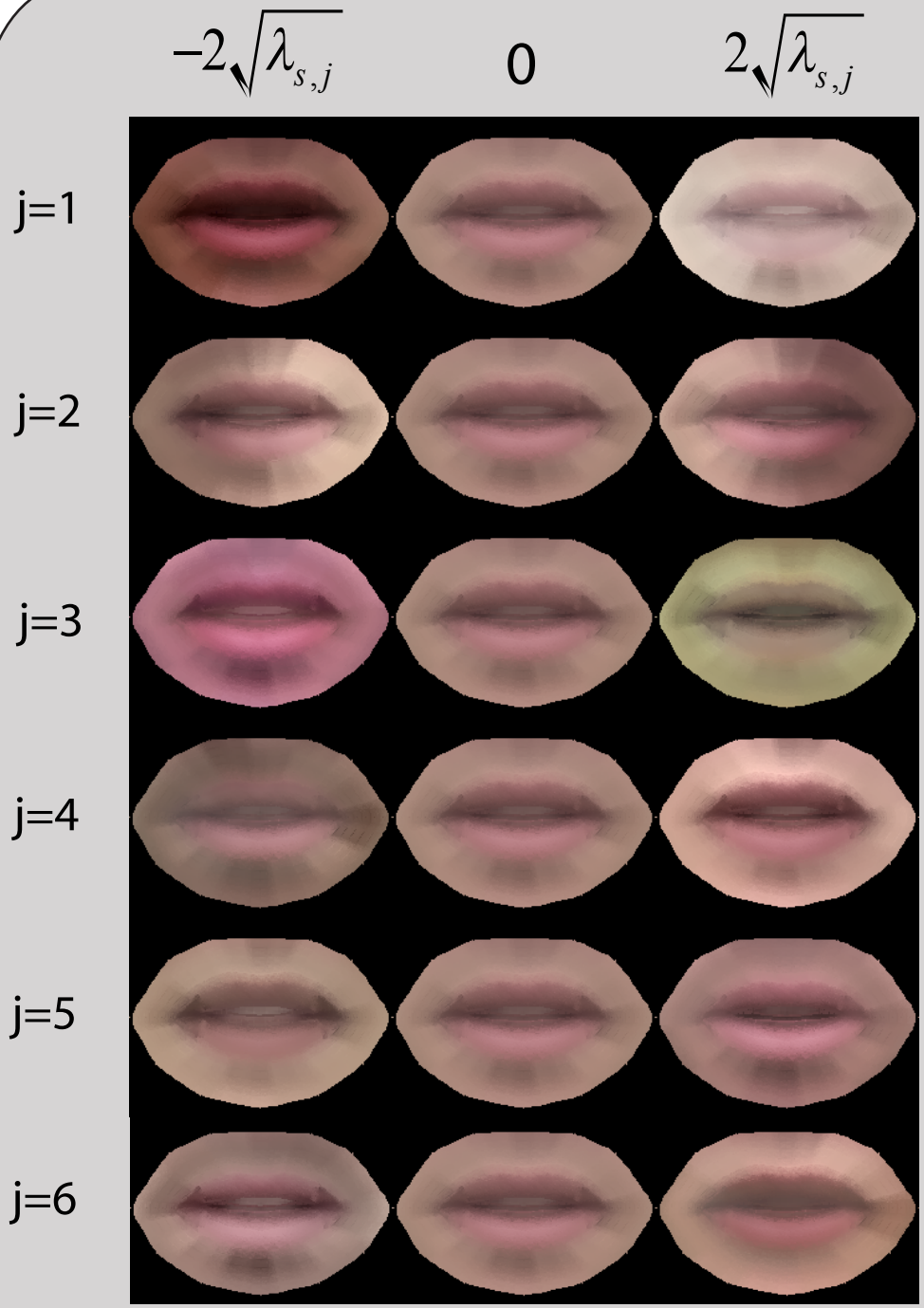


Figure 3.12 :
Les 6 modes propres du modèle actif d'apparence statique.
De haut en bas : indice du mode propre, de gauche à droite : valeur à -2*(écart-type), valeur moyenne, valeur à +2*(écart-type).
L'apparence échantillonnée est interpolée, la forme et l'apparence dynamique étant constantes et égales à leurs moyennes.

intrinsèquement dépendantes l'une de l'autre.

Il est nécessaire de normaliser les différences d'unités mais nous souhaitons aussi augmenter le poids de la forme par rapport à l'apparence dynamique puisque nous considérons que les variations de cette dernière sont causées principalement par les variations de forme. Cela est effectué par le coefficient de pondération W (le coefficient 2 ayant été déterminé empiriquement) :

$$W = 2 \times \frac{\sum_{j=1}^{nb_d} \lambda_{d,j}}{\sum_{j=1}^{nb_x} \lambda_{x,j}} \qquad \text{(eq. 3.6)}$$

avec pour nos données nb_x=8 et nb_d=40.

$\mathbf{P_c}$ est alors la matrice contenant les vecteurs propres de la matrice de covariance $\mathbf{Cov_c}$ en conservant 95% de la variance, soit 11 modes avec les valeurs propres λ_k, $1 \leq k \leq 11$). $\overline{C}$ étant nul par construction puisque les paramètres ont une distribution gaussienne de moyenne nulle.

$$\overline{\boldsymbol{C}} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{d}_i \end{pmatrix} \simeq 0,\ \mathbf{Cov_c} \simeq \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{d}_i \end{pmatrix}\begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{d}_i \end{pmatrix}^T \qquad \text{(eq. 3.7)}$$

La figure 3.13 montre la distribution des trois premiers modes propres du modèle

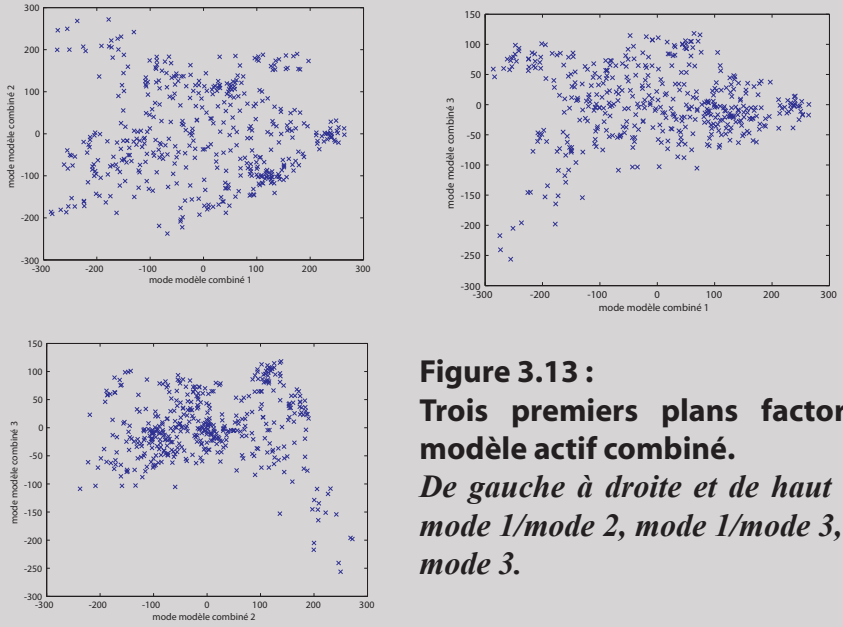


Figure 3.13 :
Trois premiers plans factoriels du modèle actif combiné.
De gauche à droite et de haut en bas : mode 1/mode 2, mode 1/mode 3, mode 2/ mode 3.

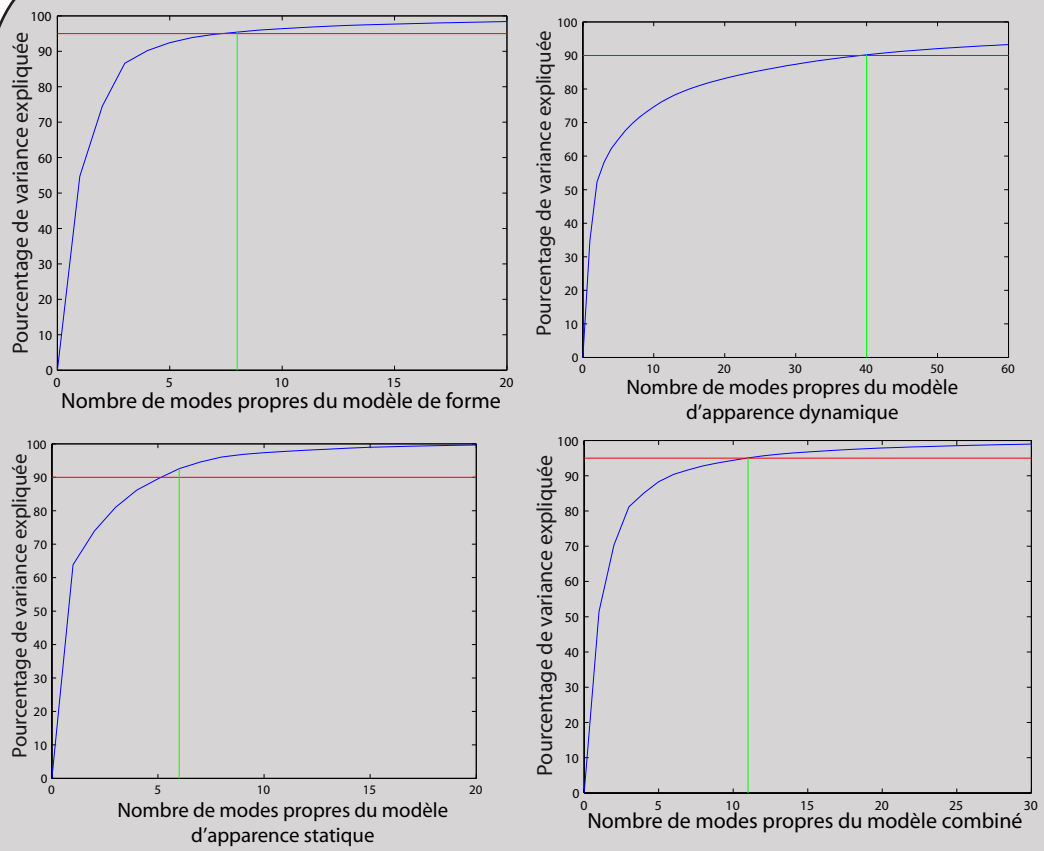


Figure 3.14 :
Evolution du nombre de paramètres des modèles de forme, d'apparence dynamique et statique et du modèle combiné en fonction du pourcentage de variance expliquée dans chaque cas.
En rouge : niveau de variance expliquée désiré, en vert : nombre de modes retenus.

combiné les uns par rapport aux autres. Leurs allures suggèrent que l'hypothèse linéaire est relativement satisfaisante avec nos données et qu'il n'est pas nécessaire de complexifier le modèle en ayant recours à des modèles non linéaires ([Cootes, 1997], [Sozou, 1994], [Sozou, 1997]) d'autant plus que les EGB imposeront des limites de variations aux modes propres. La même observation peut être faite avec l'apparence statique.

La figure 3.14 présente l'évolution de nombre de paramètres en fonction de la variance expliquée pour chacun des quatre modèles : forme, apparence dynamique, apparence statique et combiné forme/apparence dynamique.

Enfin, la figure 3.15 montre l'effet des 6 premiers modes de variation du modèle actif combiné (notons que l'on a ajouté l'apparence statique moyenne pour améliorer le rendu visuel). Le premier mode correspond au sourire et le deuxième à une protusion, les modes suivants étant moins évidents à interpréter.

Finalement toute forme et apparence échantillonnée de la base d'apprentissage, ou

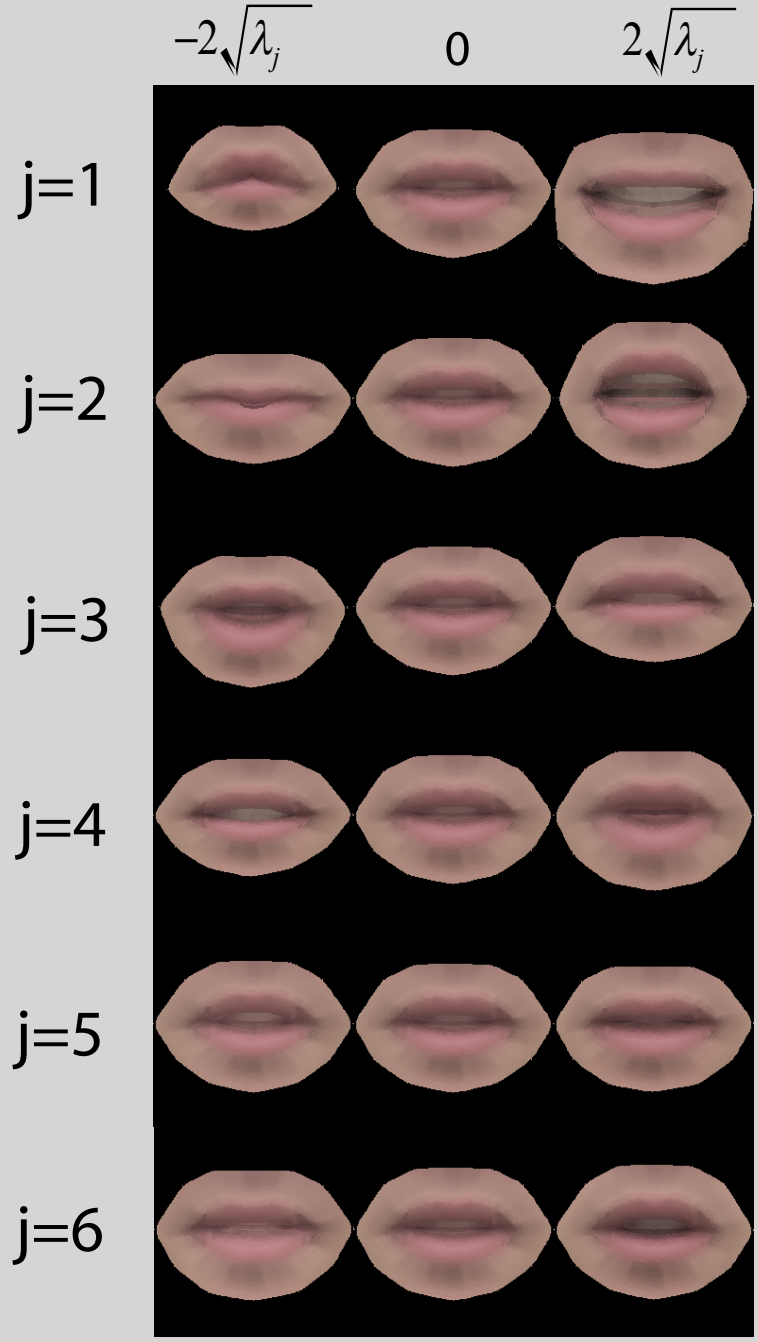


Figure 3.15 :
Les 6 premiers modes propres du modèle actif combiné forme/ apparence dynamique.
De haut en bas : indice du mode propre, de gauche à droite : valeur à -2*(écart-type), valeur moyenne, valeur à +2*(écart-type).
L'apparence échantillonnée est interpolée, l'apparence statique étant constante et égale à sa moyenne.

de nouveaux exemples, peuvent être générés en ajustant simplement $\boldsymbol{c}$ et $\boldsymbol{s}$ dans le jeu d'équation suivant :

$$(1)\ \boldsymbol{C} = \begin{bmatrix} W.\boldsymbol{x} \\ \boldsymbol{d} \end{bmatrix} = \mathbf{P}_c \boldsymbol{c} \Rightarrow \begin{cases} (2)\ \boldsymbol{X} = \overline{\boldsymbol{X}} + \mathbf{P}_x \boldsymbol{x} \\ (3)\ \boldsymbol{D} = \overline{\boldsymbol{D}} + \mathbf{P}_d \boldsymbol{d} \\ (4)\ \boldsymbol{S} = \overline{\boldsymbol{S}} + \mathbf{P}_s \boldsymbol{s} \end{cases} \Rightarrow \begin{cases} (5)\ \boldsymbol{X} = q\boldsymbol{X}_r + \boldsymbol{t} \\ (6)\ \boldsymbol{A} = \boldsymbol{D} + \boldsymbol{S} \end{cases} \quad \text{(eq. 3.8)}$$

L'équation 3.8-2 contrôle le modèle de forme, l'équation 3.8-3 contrôle l'apparence dynamique et l'équation 3.8-4 contrôle l'apparence statique tandis que l'équation 3.8-1 contrôle le modèle combinant la forme et l'apparence dynamique. Enfin les équations 3.8-5 et 3.8-6 permettent de revenir aux données : la forme non normalisée $\boldsymbol{X_r}$ et l'apparence totale $\boldsymbol{A}$.

Ainsi, segmenter la bouche sur une image inconnue consistera à trouver le meilleur jeu possible de 23 paramètres : 11 dans le vecteur de paramètre combiné $\boldsymbol{c}$, 6 dans le vecteur de paramètre de l'apparence statique $\boldsymbol{s}$ et les trois derniers contrôlant l'échelle q et la position $\boldsymbol{t}$.

3.2.3.4 Apprentissage des EGB

Enfin, les caractéristiques des vecteurs de paramètres combinés sont apprises pour chaque EGB. Les valeurs du vecteur $\boldsymbol{c}$ sont calculées pour chaque image :

$$\boldsymbol{c}_i = \mathbf{P}_c^T \begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{d}_i \end{pmatrix}, pour\ 1 \le i \le N \quad \text{(eq. 3.9)}$$

Ensuite les moyennes de chaque paramètre sont calculées pour chacun des EGB et sauvegardées dans les vecteurs $\boldsymbol{c}_j^{egb}$ (avec $1 \le j \le 4$), les limites de variations étant sauvées dans les $\boldsymbol{c}_j^{b}$ (avec $1 \le j \le 4$).

3.3 Fonctions de coût

Nous allons à présent devoir segmenter les lèvres d'un locuteur inconnu, c'est à dire optimiser à la fois les paramètres d'apparence statique et du modèle combiné forme/ apparence dynamique. Cette optimisation sera accomplie en minimisant une fonction de coût adaptée à l'algorithme DSM (voir 2.4.2).

3.3.1 Fonction de coût basée sur un calcul de champ de gradients et la valeur des pixels

Comme dans le cas mono-locuteur nous définissons une fonction de coût C_g calculée comme la somme de 6 flux de gradients adaptés aux contours (voir 2.4.3.1) et une

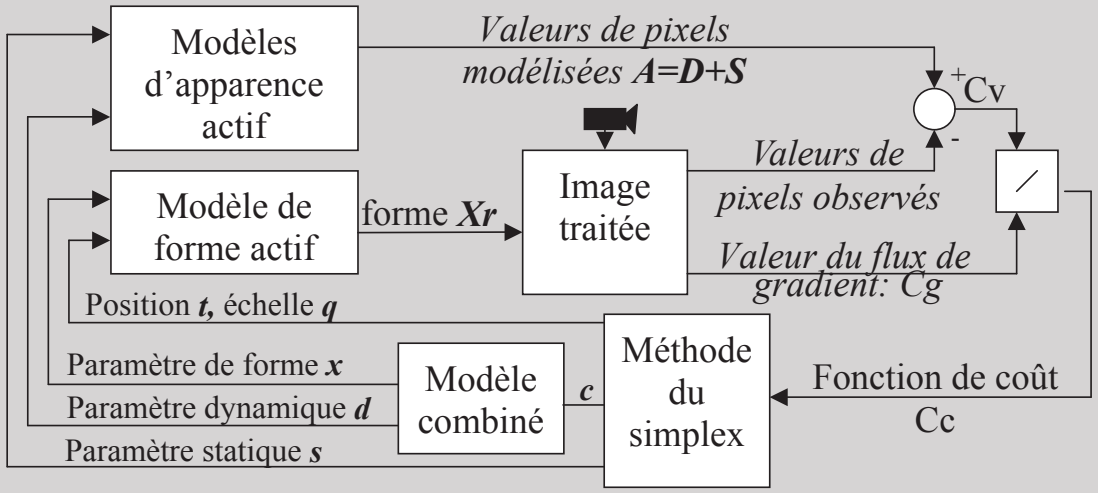


Figure 3.16 :
Segmentation de la bouche dans un cas multi-locuteurs avec la fonction de coût C_c, où C_v est une comparaison de valeurs de pixels et C_g est un flux de gradient à maximiser.

fonction C_v comparant les valeurs YCbCr de l'apparence échantillonnée modélisée et de celle observée dans l'image (voir 2.4.3.2). Enfin la fonction $C_c=C_v/C_g$ combine les critères précédents et est présentée par la figure 3.16 pour le cas multi-locuteurs. La seule différence avec le cas mono-locuteur est que l'on a rajouté l'apparence statique.

Cette fonction de coût est la première que nous avions utilisée au cours de ce travail de thèse ([Gacon, 2004], [Gacon, 2005]). En utilisant le modèle colorimétrique et les EGB pour initialiser les paramètres des modèles actifs, cette méthode donne des résultats satisfaisants dans un cas multi-locuteurs, comme le montre le tableau 3.1. La phase de suivi correspond au fait que l'apparence statique ***s*** a convergé vers un jeu de paramètres final et est donc considérée comme constante.

Ces bonnes performances nous conduiront à l'utiliser dans la suite lors de la phase de convergence du modèle d'apparence statique.

	erreur contour extérieur	erreur contour intérieur	erreur contours des dents	erreur ensemble des points	nombre d'itérations
initialisation	3.2 ± 1.7	3.6 ± 1.8	3.8 ± 2.1	3.5 ± 1.8	26
suivi	3 ± 1.4	3.2 ± 1.7	3.4 ± 2	3.1 ± 1.6	20

Tableau 3.1 :
Résultats de segmentation de la fonction de coût C_c, basée sur la valeur des pixels et les flux de gradient (images dans la base d'apprentissage).
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type.

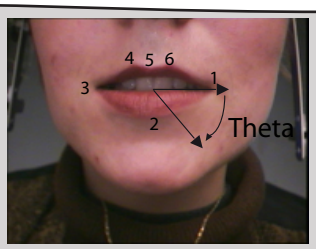
3.3.2 Fonction de coût utilisant des descripteurs locaux

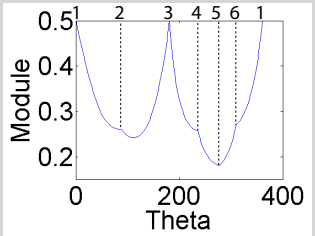
Il est apparu après expérimentation lors du chapitre 2 que la fonction de coût C_f, basée sur la prédiction de la réponse de descripteurs locaux, était parfaitement adaptée à une tâche mono-locuteur et donnait des résultats à la fois précis et robustes.

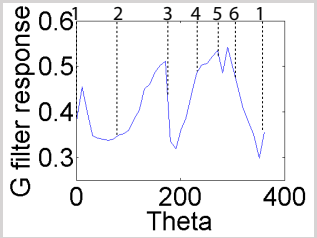
Si nous voulons adapter le principe à un cas multi-locuteurs, il faut néanmoins tenir compte de la nouvelle variabilité inter-locuteur. En effet si le locuteur devient inconnu, la réponse des filtres n'est plus uniquement déterminée par la forme : d'une personne à l'autre les couleurs de la peau et des lèvres ainsi que le contraste entre les deux est susceptible de changer. Ainsi, les paramètres de l'apparence statique, qui correspond à cette variabilité, doivent également être en entrée du réseau de neurones.

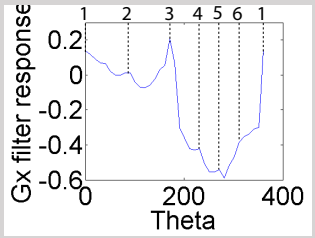
En outre, au 2.4.3.3, nous avions introduit un filtrage rétinien sur la composante Y afin de rendre la méthode moins sensible aux changements d'illuminations. Dans le cas mono-locuteur seul le modèle de forme était optimisé, dans le cas présent le fait de rajouter l'information d'apparence en entrée permet de se passer du filtrage rétinien, l'éclairage étant alors géré par le modèle d'apparence (dont les valeurs seront initialisées par le modèle colorimétrique).

Nous calculons donc pour chacune des N images de la base d'apprentissage (pour lesquelles nous connaissons la position des points de contrôle) la réponse des descripteurs en chaque localisation déterminée par la forme des contours intérieur et extérieur des lèvres sur les composantes YCbCr. La figure 3.17 montre un exemple de réponses des filtres sur Y sous la forme de courbe avec une échelle polaire (le centre étant le milieu du









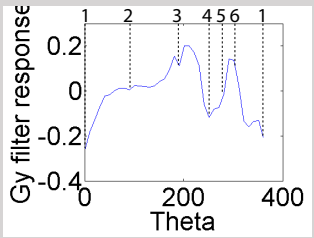


Figure 3.17 :
De gauche à droite et de haut en bas: Image traitée, contour extérieur en représentation polaire, réponses des filtres G Gx et Gy sur Y. 6 points de repère sont indiqués sur l'image et les courbes (les deux commissures, le point le plus bas et l'arc du cupidon).
Le centre des coordonnées polaires est le milieu des deux commissures. Les abscisses (angle Theta) sont en degrés.

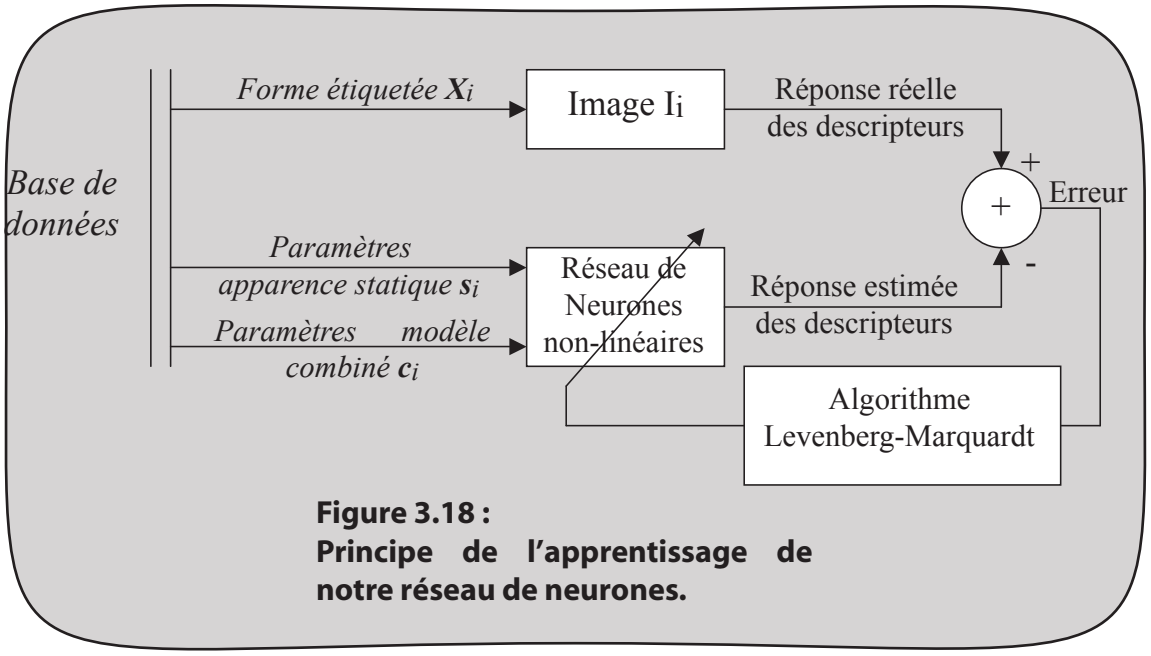


Figure 3.18 : Principe de l'apprentissage de notre réseau de neurones.

segment joignant les deux commissures). Ces réponses «idéales» des descripteurs locaux serviront donc de vérité terrain pour l'entraînement de notre réseau de neurones qui aura par conséquent 17 entrées (les paramètres $\boldsymbol{c}$ et $\boldsymbol{s}$).

On calcule donc les paramètres des modèles d'apparence statique et combiné correspondant à chaque image :

$$\boldsymbol{c}_i = \mathbf{P}_{\mathbf{c}}^{\,T} \begin{pmatrix} W.\boldsymbol{x}_i \\ \boldsymbol{d}_i \end{pmatrix}, \quad \boldsymbol{s}_i = \mathbf{P}_{\mathbf{s}}^{\,T} \left(\boldsymbol{S}_i - \overline{\boldsymbol{S}} \right) \text{ pour } 1 \le i \le N \qquad \text{(eq. 3.10)}$$

Puis le réseau de neurones est entraîné par l'algorithme de Levenberg-Marquardt selon le schéma de la figure 3.18. En procédant à une ACP afin de réduire la taille de l'espace de sortie, nous parvenons à passer d'un espace de dimension 180 (20 points d'intérêt, 3 plans et 3 filtres) à un espace de dimension 15.

De façon analogue au cas mono-locuteur, notre associateur non-linéaire de la forme et des descripteurs d'apparence sera constitué de 15 réseaux de neurones à 17 entrées (correspondant aux paramètres des modèles actifs) et une sortie. Après une méthode par essai-erreur (et par analogie au modèle locuteur), nous avons déterminé un nombre adapté d'unités cachées égal à 8, soit le nombre de paramètres du modèle de forme (voir figure 3.19).

Pour chaque réseau, nous avons 17x8 paramètres pour la première couche et 8 pour la seconde, soit au total 144 coefficients à déterminer. Notre vérité terrain étant constituée de 900 images, nous avons un rapport de 6.25 entre le nombre de paramètres à déterminer et le nombre de données ce qui est au-dessus de la limite de 5 généralement admise. Même si le risque de surapprentissage est relativement limité, il serait néanmoins préférable d'augmenter la taille de la base d'apprentissage dans l'idéal.

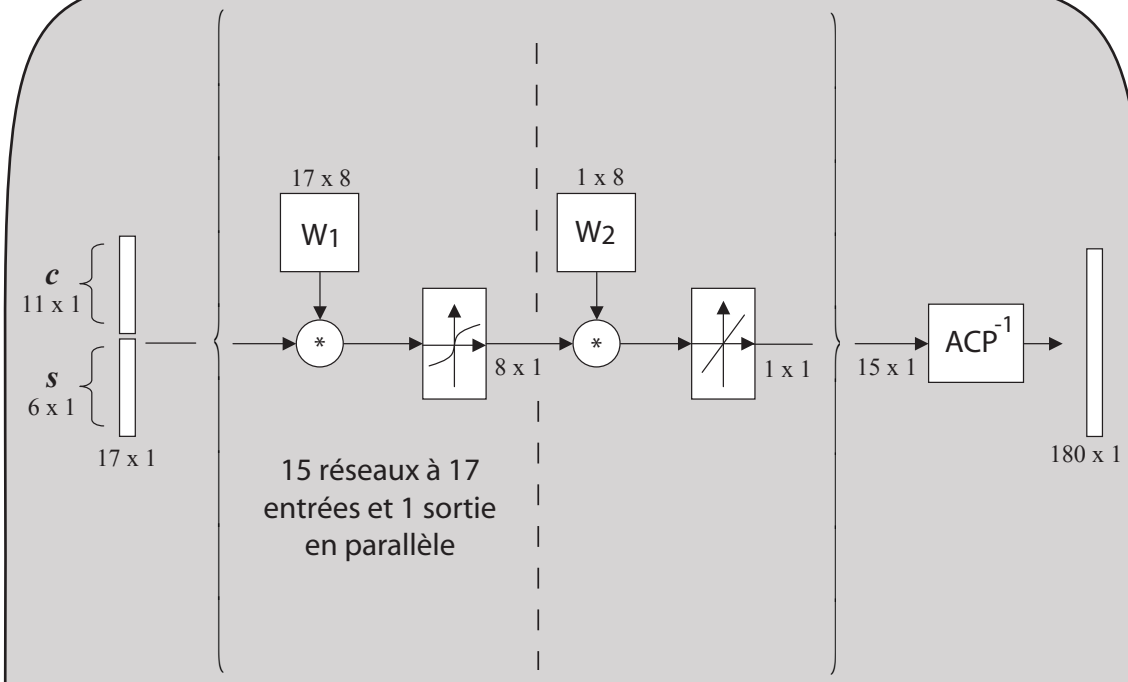


Figure 3.19 :
Structure de notre associateur non-linéaire.
Il s'agit d'une série de 15 réseaux de neurones non-linéaires à deux couches prédisant l'apparence locale des 20 points de lèvres à partir des paramètres de formes. Les entrées des réseaux sont d'abord multipliées par la matrice W_1 des poids de la première couche, puis chaque poids obtenu est traité par une fonction tangente-hyperbolique. Puis, la sortie de la première couche est multipliée par la matrice de poids de la seconde couche qui est ensuite traitée par une fonction linéaire. On obtient ainsi les 15 sorties des réseaux qui sont projetées dans l'espace des réponses des descripteurs locaux par ACP inverse.

3.3.3 Fonction de coût d'initialisation

Le tableau 3.2 montre que si la fonction de coût basée sur les descripteurs locaux C_f a donné d'excellents résultats dans le chapitre 2, elle ne devient réellement performante par rapport à C_c (combinaison d'un critère pixel et gradient) en multi-locuteurs qu'à partir du moment où l'on peut considérer que le problème est devenu mono-locuteur, c'est à dire quand l'apparence statique a convergé.

On définit une fonction de coût d'initialisation C_i qui sera la somme pondérée des fonctions de coût C_c et C_f. Elle est mise en oeuvre sur les premières images d'une séquence, lorsque l'apparence statique est encore relativement mal connue. Le tableau 3.2 montre que l'on obtient un gain important par rapport aux fonctions de coût calculées séparément. Une fois ***s*** connue, le critère C_f devient plus rapide et même légèrement plus précis que C_i.

	erreur initialisation	erreur suivi	nombre d'itérations initialisation	nombre d'itérations suivi
C_c	3.5 ± 1.8	3.1 ± 1.6	26	20
C_f	3.6 ± 2.1	2.8 ± 1.3	17	12
C_i	3 ± 1.4	2.9 ± 1.3	22	16

Tableau 3.2 :
Résultats de segmentation pour la fonction de coût C_c (valeur des pixels et flux de gradient), la fonction de coût C_f (réponse de filtres gaussiens), et la fonction d'initialisation C_i qui combine les deux autres (images dans la base d'apprentissage).
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type. Notons que C_i donne les meilleurs résultats en initialisation, tandis que C_f est la plus performante en suivi.

C_c sera donc utilisée conjointement à C_f le temps que les paramètres de $\boldsymbol{s}$ aient convergé sur une séquence d'images ce qui s'obtient généralement en 5 à 6 images. Dès que cette convergence est effective et que $\boldsymbol{s}$ est donc connu, le problème revient alors intrinsèquement à un problème mono-locuteur et dès lors, seul C_f sera utilisée pour optimiser $\boldsymbol{c}$ sur les images ultérieures de la séquence.

3.4 Implémentation pratique de la méthode de segmentation de la bouche

Nous voulons obtenir le meilleur jeu de paramètres afin que notre modèle opère la meilleure segmentation possible de la bouche.

Dans la suite nous noterons $C(I_n)(\boldsymbol{c},\boldsymbol{s},q,\boldsymbol{t})$ la valeur d'une fonction de coût appliquée à l'image I_n pour les vecteurs de paramètres $\boldsymbol{c}$ et $\boldsymbol{s}$, et pour l'échelle q et la position $\boldsymbol{t}$.

Afin d'améliorer la vitesse de convergence et de diminuer le risque d'obtenir un minimum local, nous allons définir la meilleure initialisation possible pour l'algorithme ainsi que des intervalles de recherche appropriés pour les paramètres.

L'initialisation classique du DSM est la valeur moyenne des paramètres avec une amplitude des intervalles de recherche correspondant à 3 fois l'écart-type des variations de chaque paramètre. L'optimisation de ces choix augmente la précision et la robustesse et diminue le nombre d'itérations nécessaires.

Dans notre méthode, pour la première image d'une séquence, les paramètres du Simplex sont initialisés :

- pour $\boldsymbol{c}$: en testant les jeux de paramètres moyens correspondant aux différents EGB et en choisissant celui qui donne le coût minimal
- pour $\boldsymbol{s}$: en utilisant la classification de pixel
- pour q et $\boldsymbol{t}$: en utilisant le modèle de commissures.

Première image I_1 d'une séquence

Le modèle de commissures est utilisé afin de détecter les deux points clés. Cela donne des valeurs initiales généralement fiables pour la position de la bouche $\boldsymbol{t}$ ainsi que pour son échelle q.

Le modèle colorimétrique est ensuite utilisé afin de classifier grossièrement les pixels de peau et de lèvres sur la première image I_1 ce qui permet un début de connaissance sur les caractéristiques du locuteur.

A partir de cette classification on obtient donc les valeurs moyennes et écarts-types des composantes YCbCr pour les lèvres et pour chaque quadrant de la zone de peau (voir figure 3.7). Ces valeurs sont utilisées afin d'obtenir par régression linéaire un jeu de valeur initial $\boldsymbol{s}_0$ pour $\boldsymbol{s}$ (représentatif de la chromacité du locuteur et de la luminance générale de la scène) qui sera l'apparence statique initiale.

Afin d'obtenir une estimation initiale de $\boldsymbol{c}$, nous testons l'Etat Général de la Bouche (EGB). Pour cela on calcule pour chaque EGB: $C_c(I_1)(\boldsymbol{c}_j^{egb},\boldsymbol{s}_0,q,\boldsymbol{t})$ pour $1 \leq j \leq 4$, les $\boldsymbol{c}_j^{egb}$ étant les jeux de paramètres moyens calculés lors de l'apprentissage. Le minimum donne l'initialisation du DSM : $(\boldsymbol{c}_m^{egb},\boldsymbol{s}_0,q,\boldsymbol{t})$.

Nous procédons ensuite à la minimisation de $C_i(I_1)(\boldsymbol{c},\boldsymbol{s},q,\boldsymbol{t})$ par DSM et nous trouvons le jeu de paramètre final : $(\boldsymbol{c}_1,\boldsymbol{s}_1,q_1,\boldsymbol{t}_1)$.

Les intervalles de recherches pour le vecteur de paramètres $\boldsymbol{c}$ sont définis par l'EGB et les limites de paramètres $\boldsymbol{x}_{jm}^{b}$ (les $\boldsymbol{x}_j^{b}$ ayant été calculées lors de l'apprentissage (3.3.2.4)).

Les intervalles de recherche pour le vecteur $\boldsymbol{s}$ sont $3\sqrt{\lambda_{s,k}}$.

Le DSM est considéré comme ayant convergé quand la différence entre les valeurs maximum et minimum du Simplex est en-dessous d'un seuil et que la valeur du minimum est également en-dessous d'un seuil fixé empiriquement.

Suivi tant que $\boldsymbol{s}$ n'a pas convergé

Nous calculons tout d'abord un critère afin de déterminer si la nouvelle image a beaucoup changé par rapport à la précédente. Pour l'image I_{n+1}, il s'écrit :

$$\left| \frac{C_c\left(I_{n+1}\right)\left(\boldsymbol{c}_n,\boldsymbol{s}_n,q_n,\boldsymbol{t}_n\right) - C_c\left(I_n\right)\left(\boldsymbol{c}_n,\boldsymbol{s}_n,q_n,\boldsymbol{t}_n\right)}{C_c\left(I_n\right)\left(\boldsymbol{c}_n,\boldsymbol{s}_n,q_n,\boldsymbol{t}_n\right)} \right| \leq 20\% \qquad \text{(eq. 3.11)}$$

Si ce critère est vérifié, cela signifie que la bouche a peu bougé par rapport à l'image précédente. Nous minimiserons alors $C_i(I_{n+1})(\boldsymbol{c},\boldsymbol{s},q,\boldsymbol{t})$ par DSM, avec $\boldsymbol{c}_n$ comme estimation initiale et des intervalles de recherche réduits correspondant à $0{,}5\sqrt{\lambda_k}$ pour le vecteur de paramètres $\boldsymbol{c}$ et q_n et $\boldsymbol{t}_n$ comme estimations initiales de q et $\boldsymbol{t}$.

Si le critère est négatif, nous testons à nouveau les EGB et détectons à nouveau les commissures.

Pour le vecteur de paramètre $\boldsymbol{s}$, l'estimation initiale est la moyenne de tous les $\boldsymbol{s}p$ précédents ($1 \leq p \leq n$), cette moyenne étant pondérée par les valeurs respectives

Image Initiale I_1 de la séquence

Paramètres	Point Initial Simplex	Intervalle recherche
Paramètres du modèle combiné $\boldsymbol{c}$	Après détermination de l'EGB jm : c_j^{egb}	Après détermination de l'EGB j : c_m^{egb}
Paramètres d'apparence statique $\boldsymbol{s}$	Donné par la classification de pixels	Pour chaque paramètre k : $3\sqrt{\lambda_{s,k}}$
Echelle q Position $\boldsymbol{t}$	Déterminé par la détection des commissures	Un dixième de la largeur de la zone du visage

Phase de Suivi pour l'image I_n, $\boldsymbol{s}$ n'a pas encore convergé (cas multi-locuteur)

Paramètres	Point Initial Simplex	Intervalle recherche
Paramètres du modèle combiné $\boldsymbol{c}$	Valeurs image précédente	Pour chaque paramètre k : $\sqrt{\lambda_k}$
Paramètres d'apparence statique $\boldsymbol{s}$	Moyenne pondérée des valeurs des images précédentes	Pour chaque paramètre k : $3\sqrt{\lambda_{s,k}}/n$
Echelle q Position $\boldsymbol{t}$	Valeurs image précédente	Un dixième de la largeur de la zone du visage

Phase de Suivi pour l'image I_n, $\boldsymbol{s}$ a convergé (cas mono-locuteur)

Paramètres	Point Initial Simplex	Intervalle recherche
Paramètres du modèle combiné $\boldsymbol{c}$	Valeurs image précédente	Pour chaque paramètre k : $\sqrt{\lambda_k}$
Paramètres d'apparence statique $\boldsymbol{s}$	Valeurs fixes $\boldsymbol{sf}$: résultat de la convergence de $\boldsymbol{s}$	Plus d'optimisation
Echelle q Position $\boldsymbol{t}$	Valeurs image précédente	Un dixième de la largeur de la zone du visage

Tableau 3.3 :
Initialisation de l'algorithme DSM selon les cas de figure.

$C_c(\boldsymbol{c}_p,\boldsymbol{s}_p,q_p,\boldsymbol{t}_p)$, ceci afin d'accorder plus de poids aux paramètres qui ont donné les fonctions de coût les plus faibles et donc les meilleures segmentations. Les intervalles de recherche pour chaque paramètre sont $3\sqrt{\lambda_{s,k}}/(1+n)$.

Enfin si $C_i(I_{n+1})(\boldsymbol{c}_{n+1},\boldsymbol{s}_{n+1},q_{n+1},\boldsymbol{t}_{n+1}) > 2C_i(I_n)(\boldsymbol{c}_n,\boldsymbol{s}_n,q_n,\boldsymbol{t}_n)$, nous supposons que le modèle a divergé. Cela peut être causé par un brusque changement d'éclairage par exemple. L'image suivante sera alors traitée comme une image initiale.

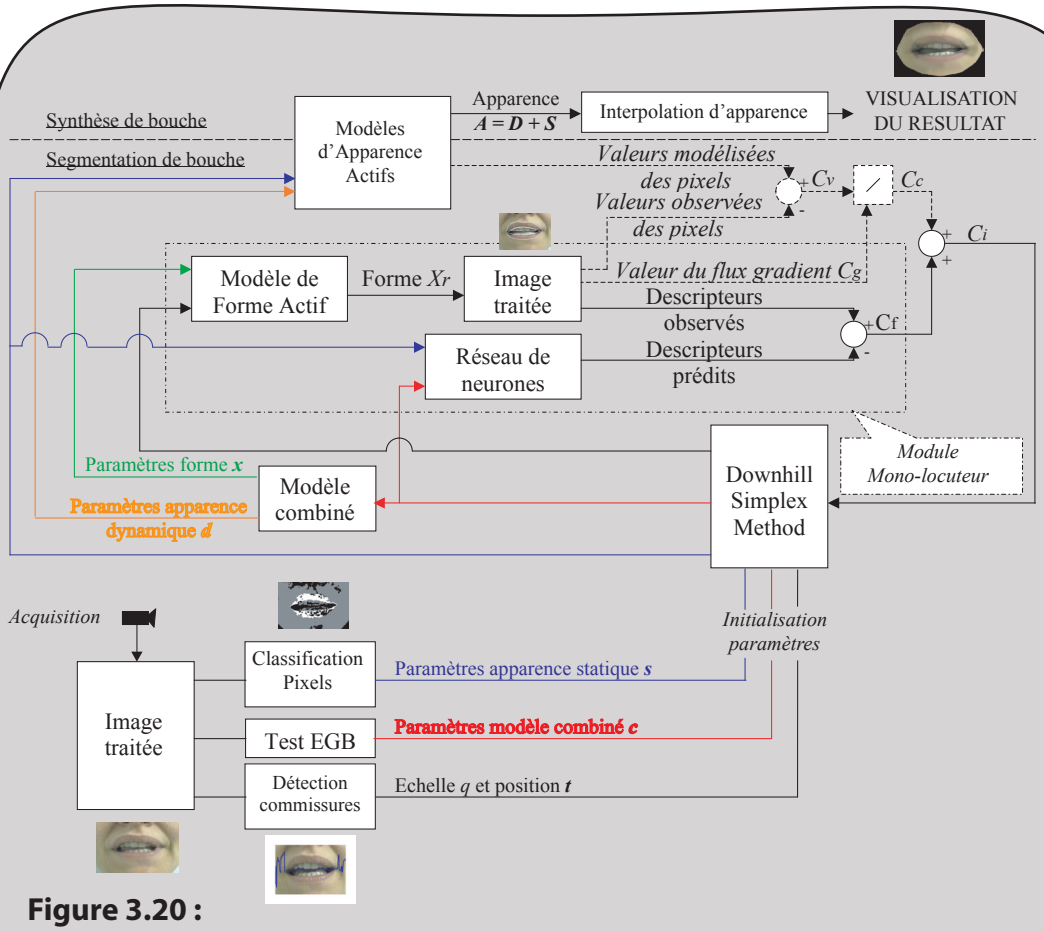


Figure 3.20 :
Méthode de segmentation multi-locuteurs, phase de convergence de l'apparence statique.
Les valeurs modélisées et observées des pixels (en italique sur les fonctions en pointillé) sont comparées tant que s n'a pas encore convergé, ce qui correspond à la fonction de coût d'initialisation C_i.
Les modèles d'apparence sont utilisés pour interpoler une bouche synthétique.

*Suivi quand **s** a convergé*

s est réputé avoir convergé quand la différence entre les paramètres d'une image à l'autre est passé sous un seuil fixé. Cela arrive généralement après 5 ou 6 images.

Le jeu de paramètre final pour ***s*** adapté au locuteur est alors appelé ***sf.***

Nous calculons tout d'abord un critère afin de déterminer si la nouvelle image a beaucoup changé par rapport à la précédente. Pour l'image I_{n+1}, il s'écrit :

$$\left| \frac{C_f(I_{n+1})(\boldsymbol{c}_n, \boldsymbol{s}_f, q_n, \boldsymbol{t}_n) - C_f(I_n)(\boldsymbol{c}_n, \boldsymbol{s}_f, q_n, \boldsymbol{t}_n)}{C_f(I_n)(\boldsymbol{c}_n, \boldsymbol{s}_f, q_n, \boldsymbol{t}_n)} \right| \leq 20\% \qquad \text{(eq. 3.12)}$$

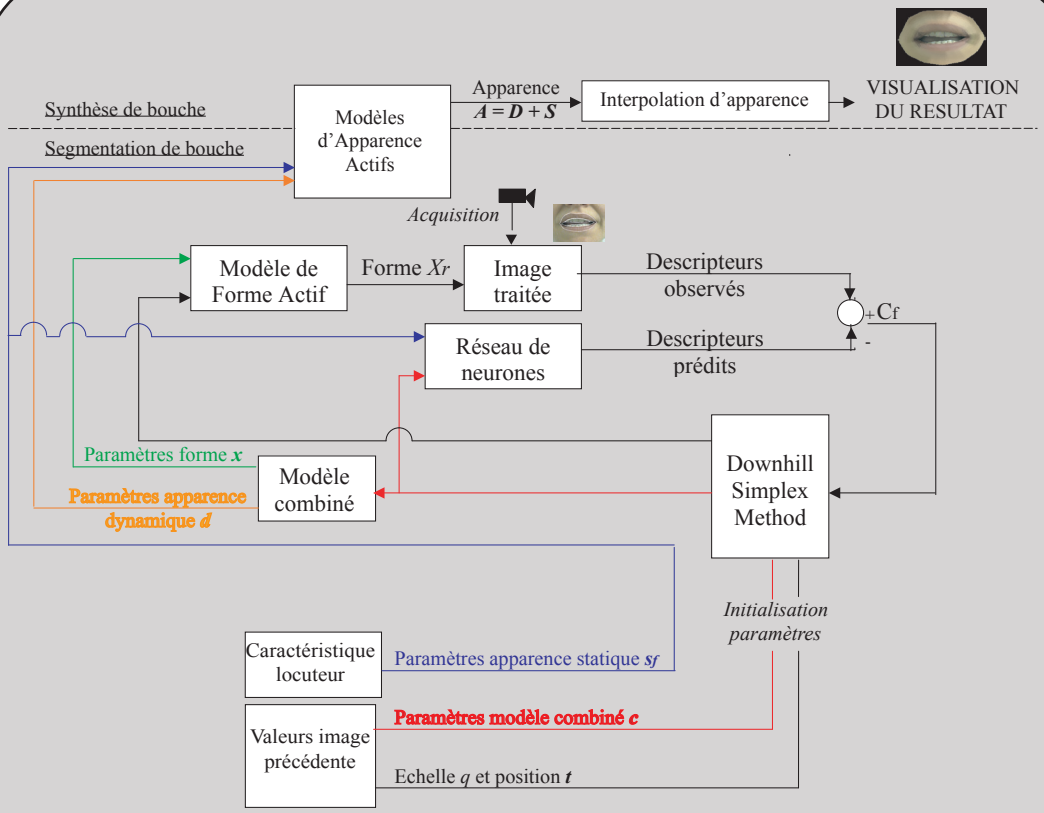


Figure 3.21 :
Méthode de segmentation multi-locuteurs, phase de suivi après convergence de l'apparence statique.
Quand s a convergé, la fonction de coût est dès lors uniquement calculée comme étant l'erreur de prédiction des descripteurs locaux gaussiens par le réseau de neurones, c'est à dire Cf.
Les modèles d'apparence ne sont dès lors utilisés que pour interpoler l'apparence échantillonnée pour la synthèse de bouche.

Si ce critère est vérifié, cela signifie que la bouche a peu bougé par rapport à l'image précédente. Nous minimiserons donc $C_f(I_{n+1})(\boldsymbol{c},\boldsymbol{sf},q,\boldsymbol{t})$ par DSM, avec $\boldsymbol{c}_n$ comme estimation initiale et des intervalles de recherche réduits correspondant à $0{,}5\sqrt{\lambda_k}$ pour le vecteur de paramètres $\boldsymbol{c}$ et q_n et $\boldsymbol{t}_n$ comme estimations initiales de q et $\boldsymbol{t}$.

Si le critère est négatif, nous testons à nouveau les EGB et détectons à nouveau les commissures.

Enfin si $C_f(I_{n+1})(\boldsymbol{c}_{n+1},\boldsymbol{sf},q_{n+1},\boldsymbol{t}_{n+1}) > 2C_f(I_n)(\boldsymbol{c}_n,\boldsymbol{sf},q_n,\boldsymbol{t}_n)$, nous supposons que le modèle a divergé. Cela peut être causé par un brusque changement d'éclairage par exemple. L'image suivante sera alors traitée comme une image initiale.

On notera enfin qu'une fois la convergence du vecteur $\boldsymbol{s}$ achevée, les modèles d'apparence actifs statique et dynamique ne sont plus directement utilisés pour la segmentation mais uniquement pour interpoler un clone synthétique du locuteur, dont une utilisation sera

faite au chapitre 4.

Toutes les initialisations du DSM sont résumées par le tableau 3.3. La figure 3.20 présente le schéma de principe de la méthode lors de la phase de convergence de $\boldsymbol{s}$ et la figure 3.21 la méthode en suivi, lorsque que le problème est redevenu de type mono-locuteur.

3.5 Bilan

Ce chapitre a présenté l'extension du cas mono-locuteur et la mise en oeuvre de la fonction de coût non-linéaire à un cadre multi-locuteurs plus large qu'au chapitre 2.

Pour cela nous avons introduit la notion d'apparence statique permettant de revenir à un problème mono-locuteur après une phase d'initialisation. Nous avons également présenté un modèle colorimétrique mis en oeuvre comme prétraitement de la méthode globale de segmentation des lèvres.

On pourra remarquer que l'approche retenue pour notre méthode de segmentation dans

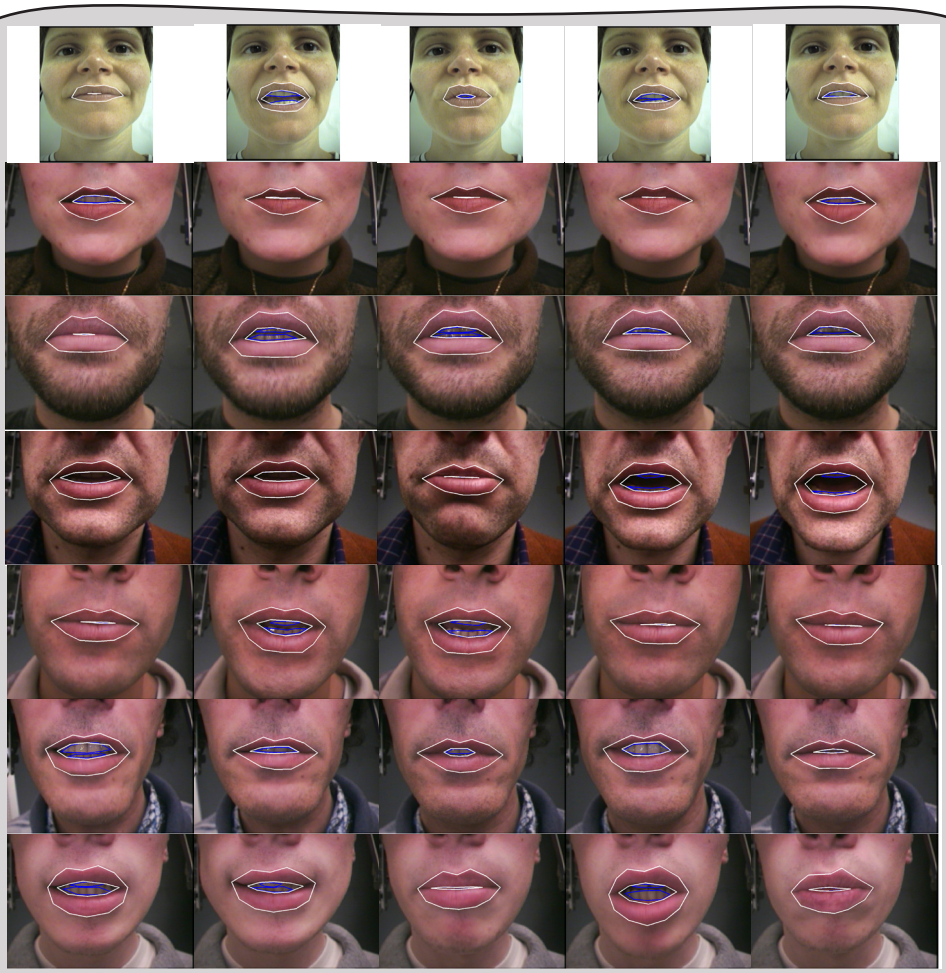


Figure 3.22-1 :
Exemples de segmentation de lèvres dans le cas multi-locuteurs.

un cas multi-locuteurs a un coût relativement élevé en opérations lors de la phase de convergence de l'apparence statique, étant donné que la fonction de coût utilisée implique trois critères différents (sur les pixels, les flux de gradient et les descripteurs locaux). Néanmoins une grande partie de cette sophistication et de cette complexité algorithmique est supprimée dès lors que l'on passe à la seconde phase, la méthode devenant alors une extension de la méthode du chapitre 2 n'impliquant plus que la prise en compte des descripteurs locaux.

La figure 3.22 donne quelques exemples de segmentation sur les images de la base de données tandis que les performances de cette méthode et l'apport des différents prétraitements seront largement discutés au chapitre 4.

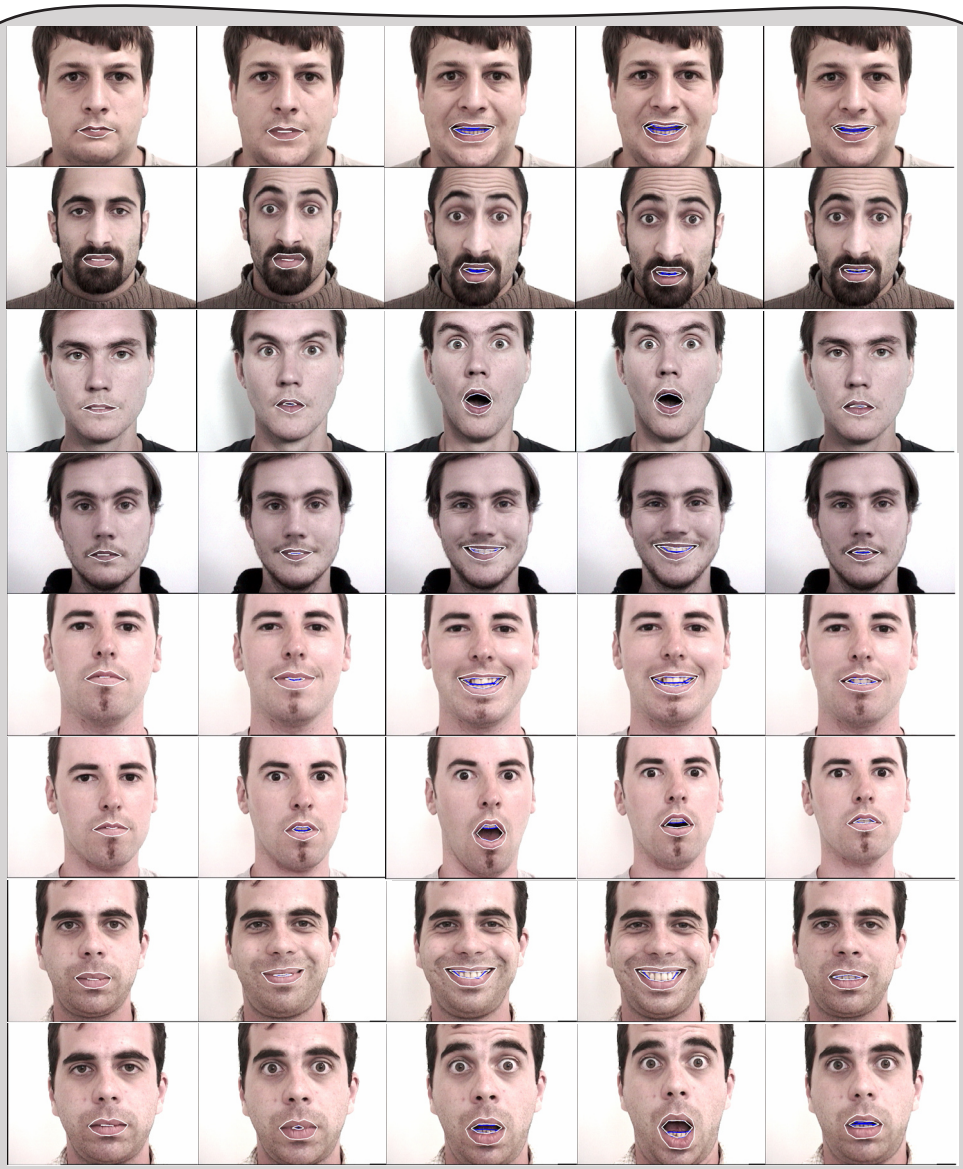


Figure 3.22-2 :
Exemples de segmentation de lèvres dans le cas multi-locuteurs.

CHAPITRE 4

Résultats et Evaluations

Ce chapitre présente tout d'abord une application de notre méthode consistant en la synthèse d'un avatar synthétique de la bouche du locuteur. Il rassemble en outre divers résultats de notre algorithme de segmentation dans le cas multi-locuteurs présenté au chapitre 3.

4.1 Création de bouches synthétiques

Après la convergence de notre modèle sur une image inconnue, nous connaissons la position des 30 points de contrôle ainsi que les valeurs YCbCr des 768 échantillons de l'apparence échantillonnée. On peut alors obtenir un clone ou avatar de la zone labiale

Figure 4.1 : Exemples de synthèses de bouche en mono et multi-locuteur. ***En haut : bouche réelle, en bas : bouche synthétique.***

Figure 4.2 :
Utilisation d'un masque pour obtenir un rendu plus réaliste de l'intérieur de la bouche.
De gauche à droite: image originale, image interpolée sans masque, puis avec masque pour la zone des dents.

du locuteur filmé en exploitant ces informations (voir figure 4.1). Les différents contours des lèvres et des dents sont modélisés par des courbes splines puis les valeurs YCbCr sont interpolées linéairement le long de ces courbes. Enfin les valeurs des autres pixels sont obtenues en effectuant une interpolation par itération triangulaire qui est une méthode d'inspiration fractale permettant d'avoir des dégradés de couleurs continus et naturels ([Carbini, 2003], [Ayoujil, 2003])

Le rendu est satisfaisant pour les lèvres et la peau mais en revanche il existe un effet de flou visuellement gênant au niveau des dents puisque les démarcations entre chacune d'entre elles ne sont pas prises en compte.

Pour avoir une représentation plus réaliste on a choisi d'utiliser des "masques" afin de remplir les zones des dents inférieures et supérieures définies par la connaissance des points de contrôle. Pour cela on a choisi une image de la base d'apprentissage dont on a sauvegardé les valeurs des zones en question. Ces deux masques seront utilisés indifféremment pour tous les locuteurs (voir figure 4.2).

Ces masques étant construits une fois pour toute, il faut néanmoins tenir compte des ombres et de l'éclairage qui peuvent être très variables sur les images traitées. Lors de la segmentation, un certain nombre d'échantillons sont placés dans les zones des dents, et une fois la convergence obtenue ils sont représentatifs des valeurs présentes sur l'image traitée. On ajuste donc simplement la moyenne des valeurs YCbCr des masques à celles des échantillons correspondant aux dents.

Bien que très perfectible (et très coûteux en calcul, l'algorithme étant non optimisé), le résultat est alors raisonnablement réaliste, suffisamment pour effectuer dans la suite des tests de compréhension et de lecture sur les lèvres.

4.2 Evaluation objective

Dans cette partie, les résultats utilisés pour évaluer notre méthode sont des erreurs de positionnement 2-D des 30 points définissant la forme obtenue lors de la segmentation d'image dont on a étiqueté les points de contrôle. Les erreurs (et leurs écart-types

méthode	pas de prétraitement	classification de pixels	recherche de l'EGB	deux prétraitements effectués
erreur	5.2 $\pm$ 1.6	3.8 $\pm$ 1.5	3.9 $\pm$ 1.5	3 $\pm$ 1.4
nombre d'itérations	38.2	32.3	27.6	22.1

Tableau 4.1 :
Améliorations apportées par les prétraitements (segmentations sur les 900 images appartenant à la base d'apprentissage).
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type.

correspondants) sont donnés en pourcentage : la différence entre les points détectés et les points annotés manuellement est normalisée par la largeur de la bouche.

4.2.1 Apport des prétraitements

Le Tableau 4.1 montre les apports en précision et en robustesse apportés par les étapes de prétraitements améliorant l'initialisation : classification des pixels et détermination de l'EGB. Comme, en pratique, ces initialisations sont uniquement mises en oeuvre sur la première image d'une séquence, toutes les images segmentées lors des tests le sont indépendamment les unes des autres.

On constate que chacun des prétraitements apporte un gain de précision de l'ordre de 20%, l'utilisation conjointe des deux prétraitements diminuant l'erreur d'environ 40%.

Le tableau 4.2 montre, quant à lui, l'apport de la détermination de l'EGB sur la segmentation d'une image en fonction de l'EGB de la dite image.

On constate que l'apport est particulièrement important pour les images correspondant

EGB	bouche ouverte	bouche fermée	bouche souriante	bouche en O
erreur sans recherche de l'EGB	2.8 $\pm$ 1.4	3.3 $\pm$ 1.5	4.6 $\pm$ 2.8	4.3 $\pm$ 1.9
erreur avec recherche de l'EGB	2.8 $\pm$ 1.2	2.9 $\pm$ 1.4	3.2 $\pm$ 1.6	3 $\pm$ 1.5

Tableau 4.2 :
Améliorations apportées par la détermination de l'EGB lors de la convergence pour chaque EGB (segmentations sur les 900 images appartenant à la base d'apprentissage, classification de pixels effectuée en prétraitement).
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type.

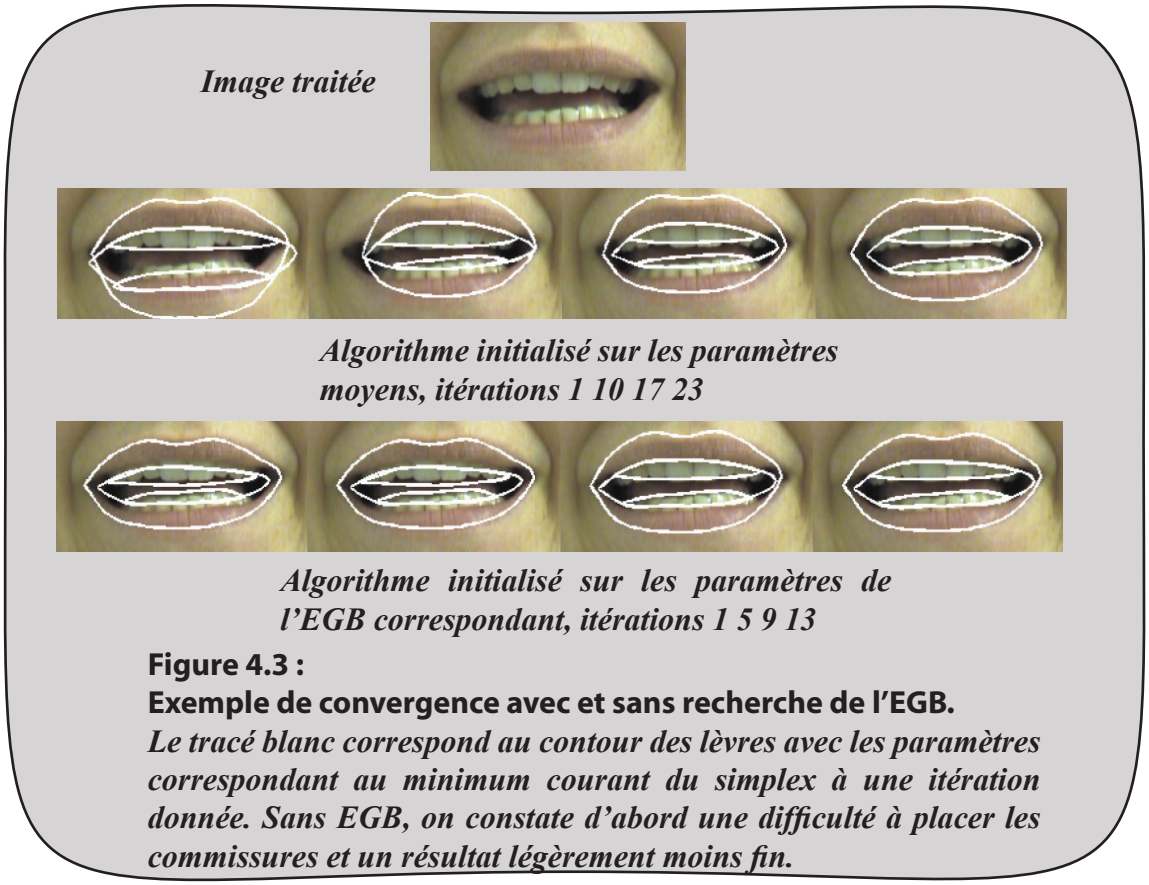


Figure 4.3 :
Exemple de convergence avec et sans recherche de l'EGB.
Le tracé blanc correspond au contour des lèvres avec les paramètres correspondant au minimum courant du simplex à une itération donnée. Sans EGB, on constate d'abord une difficulté à placer les commissures et un résultat légèrement moins fin.

aux EGB "bouche souriante" et "bouche en O". Cela découle logiquement du fait que la prise en compte de l'EGB se justifiait justement par la volonté de traiter efficacement ces cas de figures dont les paramètres correspondants s'éloignent fortement des valeurs moyennes. A l'opposé le gain sur l'EGB "bouche ouverte" est faible car cela correspond à une configuration des points proches de la forme moyenne. Enfin, pour l'EGB "bouche fermée" on a un gain correspondant au fait que l'on initialise sur une bouche déjà fermée.

La figure 4.3 montre, quant à elle, un exemple de convergence avec et sans EGB. Outre le fait que la convergence est plus rapide avec l'EGB, on constate que le contour détecté est également un peu plus fin et proche de la réalité, entre autre au niveau des commissures.

4.2.2 Résultats de segmentation pour différents protocoles de tests

Le tableau 4.3 montre les résultats de segmentation pour différents protocoles, correspondant au fait que l'image traitée est, ou non, présente dans la base d'apprentissage. Le cas le plus défavorable est le "leave-one-out" où le locuteur traité est lui-même absent de la base d'apprentissage et est donc totalement inconnu. Dans ce cas de figure, l'apprentissage a dû être refait à chaque fois que l'on testait l'algorithme sur un locuteur donné (la base d'apprentissage comptant alors moins que les N=900 images habituelles).

On constate donc que tant que le locuteur a été intégré à l'apprentissage, les résultats

position de l'erreur	contour extérieur	contour intérieur	contours des dents	ensemble des points
images présentes dans la base d'apprentissage	2.9 ± 1.3	2.9 ± 1.4	3.2 ± 1.4	3 ± 1.4
locuteur présent dans la base d'apprentissage, image absente de la base d'apprentissage	2.9 ± 1.3	3 ± 1.5	3.2 ± 1.5	3 ± 1.5
Leave-one-out : locuteur traité retiré de la base d'apprentissage	3 ± 1.5	3.3 ± 1.6	3.5 ± 1.6	3.3 ± 1.6

Image Initiale.

position de l'erreur	contour extérieur	contour intérieur	contours des dents	ensemble des points
images présentes dans la base d'apprentissage	2.7 ± 1.3	2.8 ± 1.4	3 ± 1.4	2.8 ± 1.3
locuteur présent dans la base d'apprentissage, image absente de la base d'apprentissage	2.7 ± 1.3	2.8 ± 1.5	3.1 ± 1.4	2.8 ± 1.3
Leave-one-out : locuteur traité retiré de la base d'apprentissage	3 ± 1.4	3.1 ± 1.4	3.3 ± 1.5	3.1 ± 1.3

Image en suivi.

Tableau 4.3 :
Erreurs pour différents protocoles de tests pour l'image initiale puis en suivi.
Les erreurs sont données en pourcentage de l'échelle de la bouche : erreur moyenne $\pm$ écart type.

sont similaires, que l'image soit présente ou absente de la base de données. Pour le leave-one-out, la dégradation des performances est déjà plus marquée, même si la précision des résultats reste dans le même ordre de grandeur.

Il faut néanmoins nuancer ces résultats par le fait que notre apprentissage n'a été effectué qu'avec un nombre relativement réduit d'images et ne peut prétendre représenter l'ensemble des possibilités pouvant être rencontrées.

En effet, si notre modèle de forme peut sans doute espérer s'adapter à des cas très variés, notre modèle d'apparence ne pourra pas gérer des cas où la distribution des valeurs YCbCr serait trop différente de celles présentes dans notre apprentissage. Une image acquise avec une caméra différente aurait donc une forte probabilité de mettre en échec

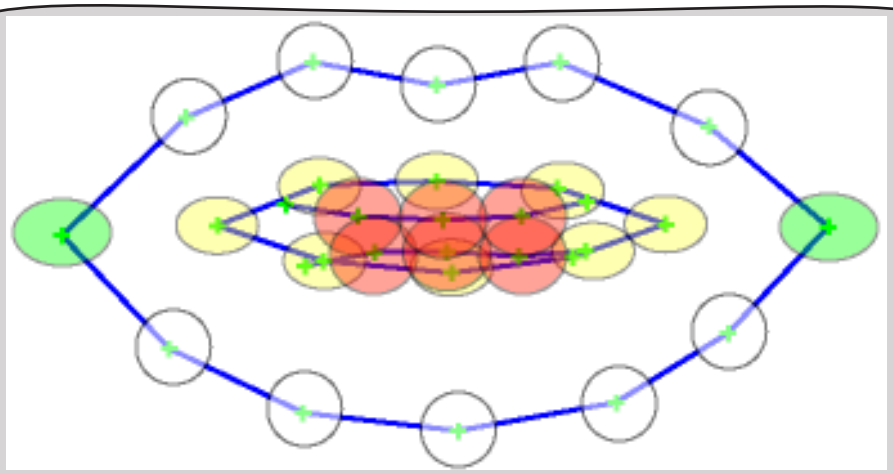


Figure 4.4 :
Illustration de l'incertitude de positionnement des points de contrôle.
En vert : commissures, blanc : contour extérieur, jaune : contour intérieur, rouge : dents. Les ellipses symbolisent l'incertitude : leurs axes sont égaux à l'erreur moyenne + 2 fois l'écart-type dans une direction, ce qui, statistiquement, correspondra donc à 95% des cas.

nos méthodes où l'apparence est modélisée. Ainsi (et ce paradoxalement, compte tenu de ses performances inférieures dans le cadre de notre étude), la fonction de coût utilisant une simple optimisation de flux de gradient serait celle qui pourrait le mieux se généraliser à des exemples extérieurs à notre base d'image initiale. Nous discuterons par la suite d'un moyen de diminuer la dépendance de notre méthode aux conditions dans lesquelles l'image à traiter a été obtenue.

La figure 4.4 donne une illustration visuelle des erreurs de positionnement pour chaque point de la forme. L'approche par modèle blanchit la distribution des erreurs qui est presque également répartie sur les différents contours.

4.3 Evaluation subjective

4.3.1 Motivation de l'évaluation subjective

Nous avons vu au §4.1 qu'une fois les valeurs optimales des paramètres connues, nous pouvions générer un avatar synthétique de la bouche du locuteur par une interpolation triangulaire. Si l'évaluation subjective de la méthode donne des résultats apparemment convaincants, il est néanmoins difficile de l'évaluer de façon qualitative. Cela peut-être accompli en évaluant l'intelligibilité des mouvements de paroles de l'avatar, jugée par un observateur humain.

Il a été démontré que l'information visuelle améliorait l'intelligibilité de la parole dans des situations acoustiques dégradées (voir chapitre 1.1.1 et la figure 1.1). Le cas extrême est celui de la lecture labiale pure, où seule l'information visuelle est utilisée pour comprendre la parole produites par un locuteur.

Ainsi, si notre ensemble analyse-synthèse est réellement performant, il devrait permettre d'améliorer la compréhension du discours par un auditeur dans des conditions bruitées.

Pour vérifier que cette amélioration était effective et la quantifier, nous avons mené une évaluation de compréhension dans le cas de numéros de téléphone ([Gacon, 2006]).

4.3.2 Tests de compréhension

Pour mener l'expérience testant l'augmentation de la compréhension, nous disposions de 40 séquences vidéos d'une seule locutrice prononçant des numéros de téléphone à 10 chiffres (séquences qui ont été utilisées dans ce rapport comme base de données de notre modèle mono-locuteur) fournies par l'ICP. Ces 40 séquences durent 8 secondes en moyenne, ce qui représente un total d'environ 8000 images. Dans la moitié des séquences la locutrice murmure les numéros tandis que son élocution est normale dans l'autre moitié.

Nous avons segmenté la bouche sur la totalité des vidéos et nous avons généré 8 types de stimuli (pour les deux types d'élocution : normale et murmurée) :

1) Audio seul avec un bruit de référence de Rapport Signal sur Bruit (RSB)= 0 dB
2) Audio seul avec un bruit fort de RSB= -18 dB
3) Vidéo naturelle (ou originale) et audio avec un bruit de référence de RSB= 0 dB
4) Vidéo naturelle et audio avec un bruit fort de RSB= -18 dB
5) Vidéo synthétique et audio avec un bruit de référence de RSB= 0 dB
6) Vidéo synthétique et audio avec un bruit fort de RSB= -18 dB
7) Vidéo naturelle seule, sans audio
8) Vidéo synthétique seule, sans audio

La figure 4.5 présente deux exemples de stimuli visuels. A partir d'une image originale, on génère une image tronquée et centrée sur la zone des lèvres qui servira pour les stimuli

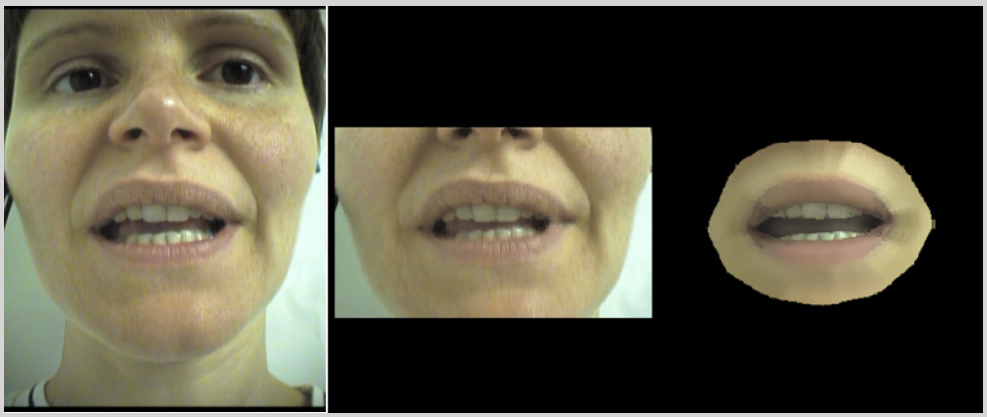


Figure 4.5 :
Exemples d'images utilisées pour notre expérience qualitative.
De gauche à droite : Image originale, exemple d'image de stimuli type "vidéo naturelle", exemple d'image de stimuli type "vidéo synthétique".

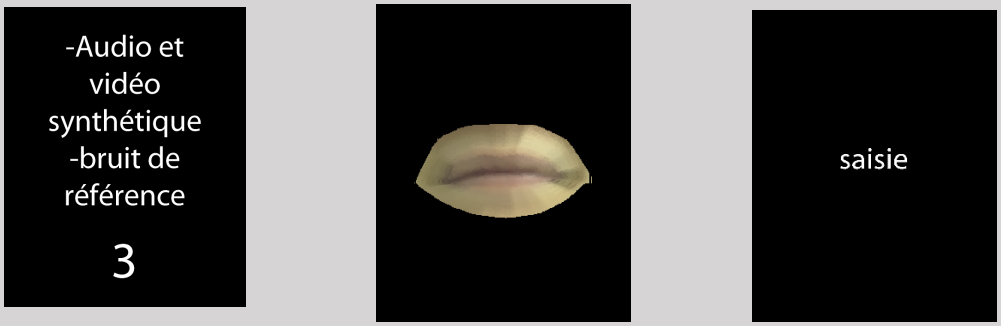


Figure 4.6 :
Protocole de l'expérience d'évaluation.
Après que le sujet a lancé la nouvelle séquence, de gauche à droite :
- Informations sur la séquence qui va être jouée avec un compte à rebours laissant le temps au sujet de se préparer à la saisie
- La vidéo est jouée, saisie des numéros de téléphone à la volée
- Saisie définitive, sans limite de temps. Une fois la tâche terminée le sujet passe à la séquence suivante.

de type "vidéo naturelle", puis on utilise notre schéma d'analyse/synthèse pour obtenir l'avatar des lèvres qui servira pour les stimuli de types "vidéo synthétique".

Le protocole de l'expérience fut le suivant :

- Chaque sujet regarda 4 séquences exemples: une en audio seule, une en audio plus vidéo naturelle, une en audio plus vidéo synthétique et une en vidéo seule, avec au moins un exemple de chaque niveau de bruit et deux exemples pour chaque type d'élocution.
- Ensuite, chaque sujet regardait 40 stimuli consécutivement (20 pour chaque type d'élocution). Comme les sujets devaient regarder l'écran pour observer le mouvement de la bouche, ils devaient prendre note de qu'ils entendaient sans regarder leur feuille. L'entrée des numéros de téléphone s'effectuait donc en deux phases : saisie à la volée puis saisie sur ordinateur une fois la séquence terminée. Ce protocole est résumé par la figure 4.6.

16 sujets ont pris part à l'expérience si bien que chacune des 40 séquences a été entendue deux fois dans chaque catégorie de stimuli. Les erreurs faites sur la retranscription des numéros de téléphones ont ensuite été comptabilisées dans chaque cas et les valeurs moyennes ainsi que les écarts types calculés. Enfin si les numéros de téléphone étaient à dix chiffres, le premier chiffre était toujours le zéro si bien qu'il n'a pas été pris en compte pour les statistiques.

4.3.3 Résultats de l'expérience

Les résultats (en pourcentage de compréhension) de nos tests de compréhension sont rassemblés dans le tableau 4.4.

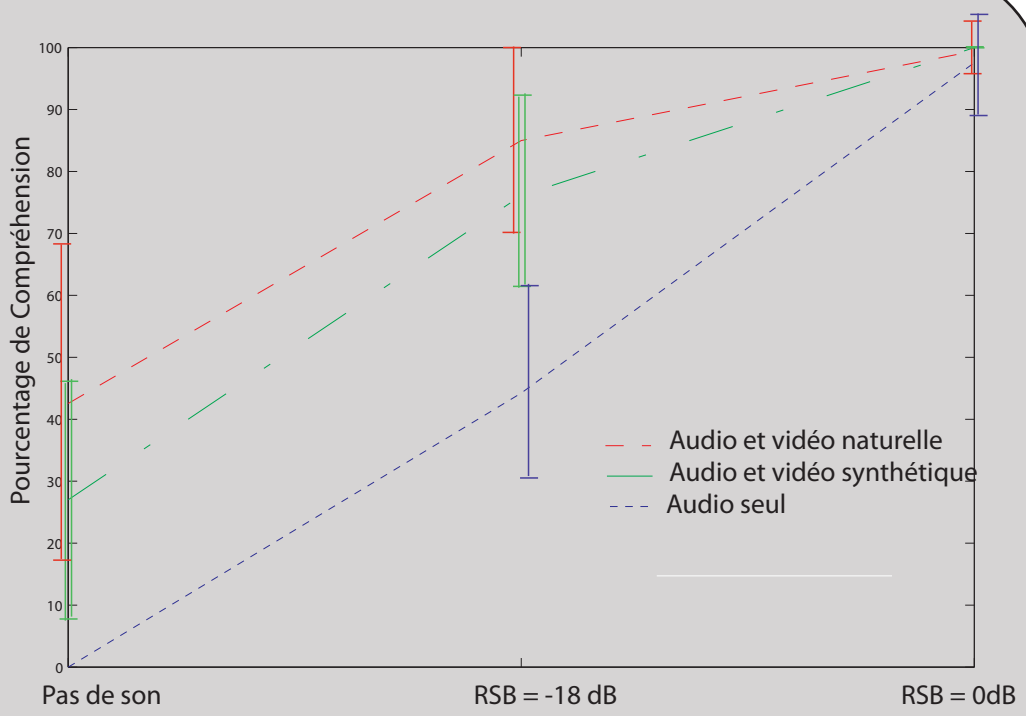
Type de Stimulus	Compréhension pour l'élocution normale	Compréhension pour l'élocution chuchotée
1) Audio seul RSB 0dB	97.5 % ± 8.4 %	92 % ± 10.3 %
2) Audio seul RSB -18dB	44.2 % ± 16.5 %	35 % ± 19.1 %
3) Audio et vidéo naturelle RSB 0dB	99.4 % ± 3.4 %	98.9 % ± 3.3 %
4) Audio et vidéo naturelle RSB -18dB	85 % ± 15.5 %	85.9 % ± 17 %
5) Audio et vidéo synthétique RSB 0dB	100 % ± 0 %	98.8 % ± 3.7 %
6) Audio et vidéo synthétique RSB -18dB	76 % ± 16 %	73.3 % ± 15.4 %
7) Vidéo naturelle seule	42.5 % ± 26 %	56.1 % ± 28.1 %
8) Vidéo synthétique seule	27 % ± 19 %	33.9 % ± 23.1

Table 4.4 :
Résultats de l'évaluation subjective.
Les valeurs sont les pourcentages moyens de bonne compréhension des chiffres et les écarts types correspondants pour chacun des 8 types de stimuli et pour chaque élocution.

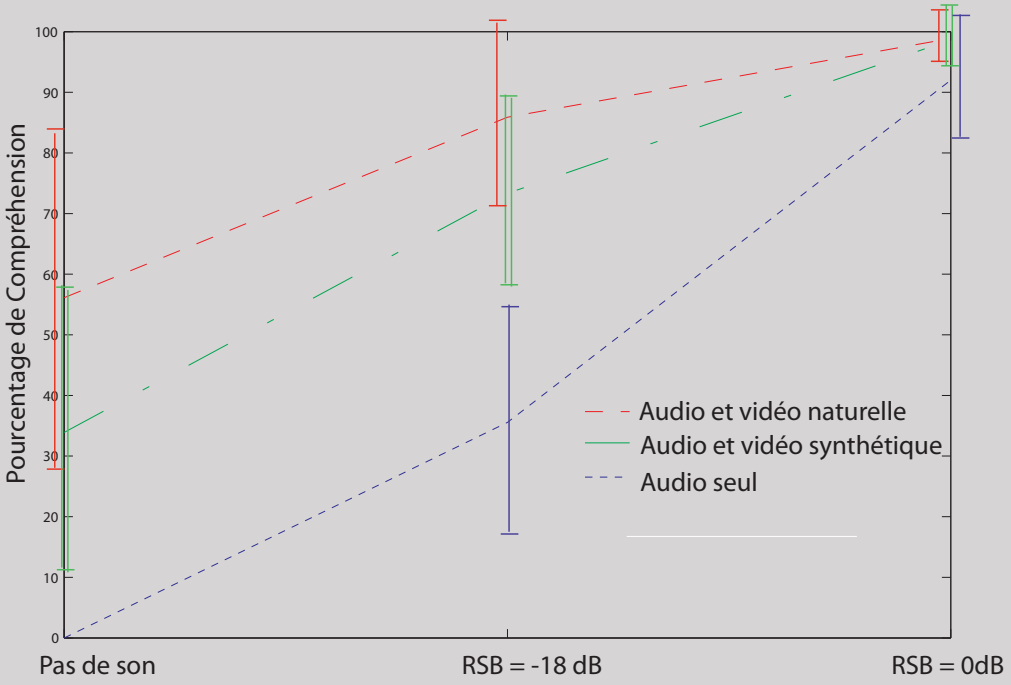
Si nous comparons les stimuli "audio et vidéo synthétique" (5-6) et les stimuli "audio seul" (1-2), nous constatons que notre synthèse permet une amélioration notable de la compréhension. Avec le bruit à -18dB, l'apport de compréhension est en effet de 32% pour l'élocution normale et 38% pour l'élocution chuchotée. On constate donc également que l'apport est supérieur en parole chuchotée (le nombre d'erreurs étant diminué de plus de moitié). Cela s'explique par le fait qu'en parole chuchotée le locuteur a une tendance à l'hyper-articulation ce qui augmente l'apport de l'information visuelle pour la compréhension. Avec le bruit de référence, on constate que les erreurs résiduelles sont pratiquement totalement supprimées par l'ajout de l'information visuelle et que les sujets ne commettent presque plus d'erreurs.

Si nous comparons les stimuli "audio et vidéo synthétique" (5-6) et les stimuli "audio et vidéo naturelle" (3-4), nous constatons que notre synthèse ne restitue pas entièrement l'amélioration de compréhension apportée par la vidéo originale. Néanmoins les résultats restent du même ordre : en moyenne, avec un bruit fort, les sujets ont fait 1 erreur de plus avec la vidéo synthétique. Avec le bruit faible, les scores de compréhension sont quasi parfaits dans les deux cas et n'apportent donc pas d'informations très pertinentes.

Si nous comparons enfin les résultats des stimuli "vidéo synthétique seule" (8) et



Elocution normale.



Elocution chuchotée.

Figure 4.7 :
Courbe de résultats de l'évaluation subjective pour chaque élocution.

“vidéo naturelle seule” (7), nous avons une différence de résultats plus désavantageuse pour notre synthèse que pour les stimuli précédents. Les sujets ont en effet effectué deux erreurs supplémentaires en lisant les lèvres du clone par rapport à celles de la locutrice originale. L’origine exacte de ces erreurs supplémentaires est difficile à déterminer mais on peut néanmoins soupçonner trois raisons: 1) une mauvaise segmentation des lèvres, 2) l’aspect synthétique et non naturel du clone perturbe davantage le sujet quand il n’a plus que cette information disponible, 3) l’absence d’indices visuels présents sur la vidéo originale (comme le mouvement de la mâchoire), seule la bouche étant clonée.

Le poids respectif de ces raisons sur la différence observée reste difficile à déterminer, néanmoins, si l’on en croit [Le Goff, 1995] (voir figure 1.1), plus le bruit augmente plus le cerveau humain va intégrer des indices visuels autres que le seul mouvement des lèvres. L’écart entre la courbe du stimuli “lèvre+audio” et celle du stimuli “visage+audio” est ainsi voisin de ce que nous observons dans notre cas sur la figure 4.7 qui reprend les résultats du tableau 4.4.

4.3.4 Tests complémentaires

Pour analyser plus finement les résultats des tests de compréhension nous avons procédé à une expérience de contrôle en testant l’amélioration de compréhension apportée par une vidéo synthétique dans le cas où les points de contrôle des contours ont été annotés manuellement et non segmentés automatiquement. Nous nous retrouvons avec une simulation de vidéo à segmentation “parfaite” (l’apparence étant extraite directement sur l’image initiale puis interpolée). De nouveaux tests d’écoute ont donc été effectués sur 5

Type de Stimuli	Compréhension de la vidéo synthétique, segmentation automatique	Compréhension de la vidéo synthétique, segmentation manuelle	Compréhension pour la vidéo naturelle
Audio et vidéo RSB 0dB	100 % $\pm$ 0 %	99.4 % $\pm$ 3.4 %	99.4 % $\pm$ 3.4 %
Audio et vidéo RSB -18dB	76 % $\pm$ 16 %	74 % $\pm$ 16 %	85 % $\pm$ 15.5 %
Vidéo seule	27 % $\pm$ 19 %	25 % $\pm$ 20 %	42.5 % $\pm$ 26 %

Tableau 4.5 :
Résultats de l’évaluation subjective pour une segmentation automatique et manuelle.
Les valeurs sont les pourcentages moyens de bonne compréhension des chiffres et les écarts types correspondant pour 3 types de stimuli en élocution normale.

vidéos en élocution normale et les résultats sont rassemblés dans le Tableau 4.5.

Si nous comparons les pourcentages de compréhension apportés respectivement par les clones générés à partir des segmentations manuelles et automatiques, nous remarquons que les valeurs sont presque identiques (la segmentation manuelle, supposée parfaite donnant même des résultats très légèrement inférieurs). La différence de compréhension observée avec la vidéo naturelle ne vient donc pas, pour la plus grande partie, d'un éventuel manque de précision de la détection (la partie analyse) mais plutôt de la génération du clone (la partie resynthèse) même s'il faut rester prudent sur l'analyse de ces résultats, l'influence respective des trois raisons envisagées pour expliquer cette différence ne pouvant être totalement certifiée.

Néanmoins, on peut supposer qu'un aspect plus réaliste du clone et une modélisation d'autres indices visuels (mouvement de la mâchoire et de la langue) pourrait améliorer les résultats de ces tests d'écoute.

4.3.5 Bilan sur l'évaluation subjective

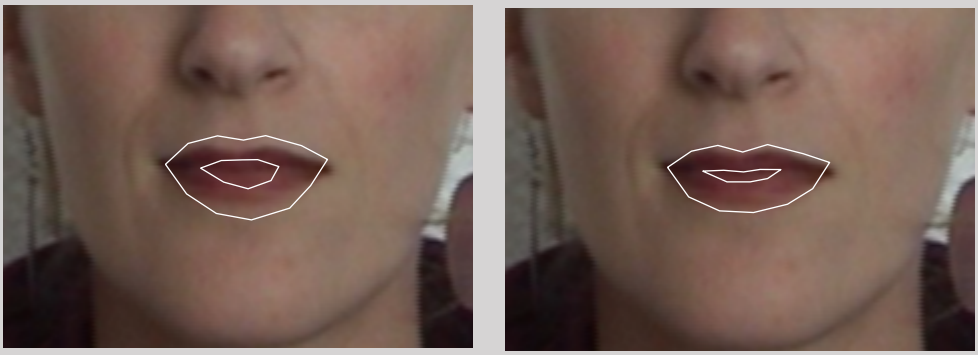
En conclusion, ces résultats montrent qu'en dépit de moins bons résultats d'intelligibilité du stimulus synthétique en lecture labiale pure vis-à-vis des mouvements naturels des lèvres, notre synthèse améliore la compréhension audiovisuelle dans un contexte bruité. En résumé, si notre synthèse ne peut pas prétendre apporter une information phonétique fiable, elle améliore le décodage acoustico-phonétique. Cela est en accord avec les modèles actuels d'intégration audiovisuelles qui mettent en avant les mécanismes de fusion où la corrélation entre les mouvements faciaux et le spectre de la parole ([Yehia, 1998]) est utilisé pour renforcer certaines zones du spectre du signal d'entrée ([Schwartz, 2004]). Ce type de filtrage adaptatif a déjà été appliqué dans le cadre de la séparation de sources audiovisuelles ([Schwartz, 2002]).

4.4 Problèmes en cas de changement de caméra

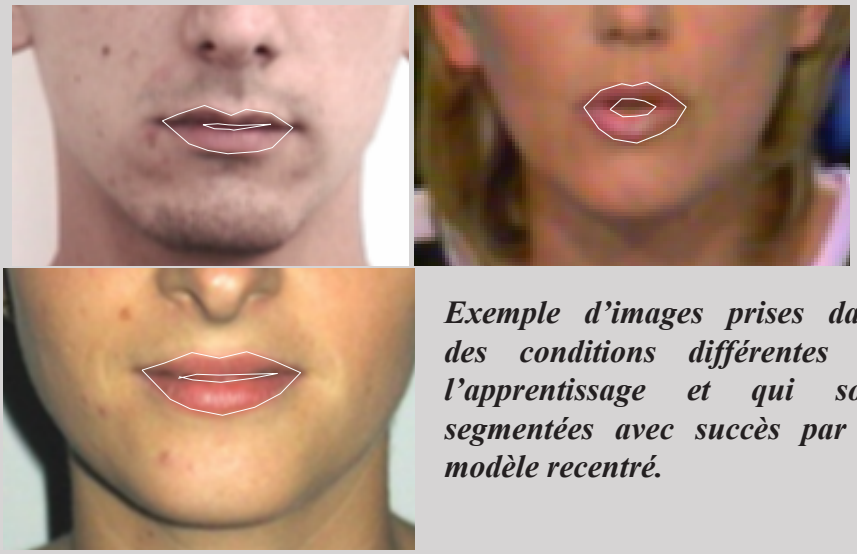
Nous avons montré que la méthode présentée dans ce rapport a de bonnes performances quand elle est appliquée à des images de locuteurs présents dans la base d'apprentissage. Les performances baissent mais restent équivalentes quand le locuteur n'est pas présent dans l'apprentissage tant que d'autres images présentes dans la base ont été prises dans des conditions similaires.

En revanche, si l'on veut appliquer la méthode à des images prises par des caméras avec des réglages colorimétriques différents de ceux présents dans la base de données, les performances vont grandement décroître, les valeurs YCbCr caractérisant le locuteur ne pouvant dans certains cas pas être modélisées car trop éloignées de ce que nous avons considéré être la "vérité terrain".

Nous sommes ici en présence de la grande limitation des modèles actifs qui ne peuvent fonctionner correctement sur des données trop différentes de celles sur lesquels ils ont été



Exemple d'image mal segmentée par le modèle normal mais bien segmentée par le modèle recentré.



Exemple d'images prises dans des conditions différentes de l'apprentissage et qui sont segmentées avec succès par le modèle recentré.

Figure 4.8 :
Prise en compte des problèmes de changement de caméra par la méthode avec données d'apparence recentrées.

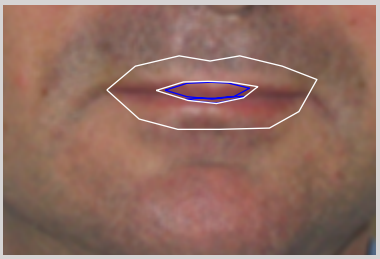
entraînés. Dans l'absolu, l'idéal serait d'entraîner le modèle d'apparence spécifiquement pour une seule et unique caméra avec des réglages qui ne changeront jamais. Néanmoins, il est possible d'améliorer la robustesse de notre modèle au changement d'acquisition des données à traiter, en appliquant l'approche évoquée au chapitre 3.2.3.1.

Nous avions en effet envisagé d'étendre le principe utilisé pour rendre le modèle colorimétrique et le modèle de commissures insensibles aux changements de caméras. Pour cela, les valeurs moyennes des composantes YCbCr pour la peau (obtenues grâce au modèle colorimétrique) sont soustraites des valeurs prises par chaque pixel. Dans ce cas, l'apparence statique ne correspondra plus qu'à la différence de teinte entre la peau et les lèvres et aux variations d'éclairement sur le visage. L'apparence dynamique reste, quant à elle, totalement inchangée de même que la méthode d'optimisation des paramètres du modèle. Le seul changement de la méthode étant le recentrage des données sur les valeurs

moyennes de la peau au moment des prétraitements.

La figure 4.8 donne un exemple d'image très mal traitée par le modèle ordinaire mais qui est bien mieux segmentée par le modèle recentré, ainsi que d'autres exemples de segmentations réussies de locuteurs absents de la base d'apprentissage et filmés dans des conditions variables.

Néanmoins, cette amélioration du modèle ne permet pas de traiter toutes les images possibles. La figure 4.9 montre en effet que dans le cas d'images trop particulières par rapport



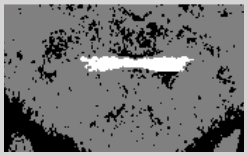


Image mal segmentée par notre méthode car trop éloignée de ce que nous avons considéré être la vérité terrain au niveau de la forme.
On constate en effet que le modèle colorimétrique des pixels effectue une bonne classification des pixels, en revanche le modèle de forme échoue à décrire les lèvres très fines du locuteur. Si la commissure gauche est bien détectée la droite est mal placée. Le contour inférieur des lèvres est presque correct mais le supérieur est très loin du réel. En outre les contours intérieurs sont très mal détectés.

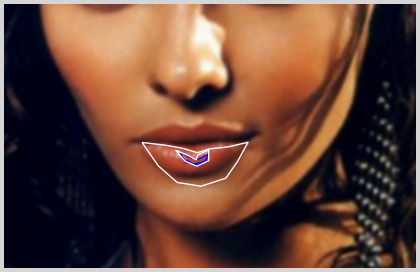


Image mal segmentée par notre méthode car trop éloignée de ce que nous avons considéré être la vérité terrain au niveau de l'apparence.
On constate que le modèle colorimétrique des pixels peine déjà à effectuer une bonne classification, un grand nombre de pixels de peau n'étant pas identifiés comme tels à proximités du contour supérieur des lèvres.
Si les commissures sont bien détectées, au final, la lèvre inférieure a été segmentée comme si elle était la bouche entière en train de sourire.

Figure 4.9 :
Exemples d'images mal traitées même par le modèle avec les données d'apparence recentrées.

aux données disponibles, le modèle ne pourra pas donner de résultats satisfaisants.

La figure 4.9 montre un exemple d'image où le locuteur a la bouche fermée et des lèvres très fines. L'absence de ce type de configuration dans notre apprentissage conduit à une segmentation très médiocre, mais ce type de problème pourrait être corrigé relativement simplement en agrandissant la base de données. En revanche, dans l'autre cas présenté, le problème vient avant tout de l'apparence, la personne étant bronzée et sous un éclairage particulier. Plutôt que d'agrandir la base d'apprentissage, un tel cas nécessiterait plutôt d'envisager un traitement des données colorimétriques.

4.5 Temps de calcul

En moyenne notre algorithme converge en 22 itérations pour une première image et en 12 itérations en suivi. Notre méthode a été uniquement implantée en langage MATLAB, seules quelques routines de calcul ayant été programmées en C++, mais sans réel souci d'optimisation.

Sur un Pentium IV, 2.5 GHz, le temps de calcul est de 1.2 sec. pour une première image et 0.6 sec. en suivi.

Si la totalité de l'algorithme était reprogrammé en C++ optimisé, on peut estimer que l'on atteindrait la vitesse nécessaire à une exécution en temps réel.

4.6 Bilan

Nous avons présenté dans cette partie une application possible de notre méthode de segmentation labiale consistant en la génération d'un avatar virtuel de la zone labiale du locuteur. L'apport en compréhension a été démontré dans le cadre d'une expérience, même s'il faut toutefois noter que, dans ce cas précis, notre modèle n'a été appliqué qu'à une seule locutrice.

En outre, l'apport des prétraitements a été validé et les performances ont été évaluées de façon quantitative dans plusieurs protocoles de tests.

Enfin, une méthode a été proposée pour élargir le champ d'application de notre schéma d'analyse/synthèse à différentes conditions d'acquisition des données, même si l'on peut estimer que la meilleure solution pour obtenir des segmentations robustes serait de calibrer le modèle spécifiquement pour le matériel qui sera par la suite utilisé pour les acquisitions.

CHAPITRE 5

Conclusion et Perspectives

5.1 Travail présenté

L'analyse labiale trouvant des applications dans de nombreux domaines relevant de l'interface homme-machine ou du multimedia, cette thèse a présenté une méthode permettant d'obtenir une segmentation précise des contours de la zone de la bouche (lèvres et dents) par une approche de modèle statistique.

Après nous être placés dans l'hypothèse que le visage a été localisé par un prétraitement adapté et qu'il est vu de face, nous nous sommes basé sur certains travaux références de la communauté scientifique et avons construit un modèle actif de forme et d'apparence de la bouche.

Afin de limiter dans un premier temps certaines difficultés dues à la variabilité des images à traiter, nous nous sommes intéressé à l'étude d'une seule locutrice filmée dans des conditions fixes.

Dans ce cadre nous avons construit pour le prélèvement de l'apparence une grille d'échantillonnage qui permet de définir précisément la localisation des pixels, ce qui est particulièrement intéressant pour l'intérieur de la bouche qui présente des non-linéarités (ouverture/fermeture, présence/absence des dents...).

Une autre description de l'apparence a été utilisée, faisant intervenir des filtres dérivés gaussiens placés sur les points de contrôle définissant la forme des contours labiaux.

Segmenter une image inconnue revient en fait à déterminer les paramètres des modèles statistiques correspondant le mieux à la réalité. Cela nécessite d'avoir une fonction de coût et une stratégie d'optimisation adaptées au problème. Pour prendre en compte le problème des minima locaux, nous avons opté pour la descente du Simplex en procédant à des étapes d'initialisation afin d'être le plus proche possible de la solution optimale avant de débuter la convergence proprement dite, ce qui améliore vitesse et robustesse.

Nous avons alors testé diverses fonctions de coût afin de déterminer la meilleure approche pour optimiser les paramètres de nos modèles statistiques. Une première fonction combine une maximisation de flux de gradient à travers les courbes ainsi que la différence entre l'apparence échantillonnée observée et celle modélisée. La seconde fait intervenir la seconde description de l'apparence et correspond à la différence entre la réponse des descripteurs locaux sur l'image traitée et la prédiction de cette réponse effectuée par un réseau de neurones non-linéaires à partir des paramètres du modèle.

Il a été montré dans ce cadre mono-locuteur que la seconde fonction était à la fois plus robuste et précise tout en nécessitant moins d'itérations pour arriver au résultat. Associée à un modèle local des commissures qui permet de placer le modèle de bouche sur l'image, cette approche non-linéaire permet de traiter le problème avec efficacité en étant en outre bien adaptée aux problèmes se posant à l'intérieur de la bouche.

Après cette première étape, la méthode a été modifiée afin d'être adaptée à un contexte impliquant différents locuteurs et des conditions d'acquisition des données variables, la philosophie retenue étant de tenter de ramener le problème à un cas mono-locuteur. Pour se faire, le modèle devait être capable de s'adapter progressivement au locuteur au fil des images d'une séquence vidéo

Dans cette optique d'adaptabilité, l'apparence échantillonnée a été séparée en deux composantes distinctes : l'apparence statique (caractéristique d'une personne et de l'éclairage de la scène) et l'apparence dynamique (qui correspond aux variations induites par le mouvement des lèvres dû, entre autre, à la parole).

Sur une séquence d'image, une fonction de coût d'initialisation (combinant les deux fonctions introduites précédemment) est alors utilisée pour déterminer l'apparence statique. Une fois celle-ci connue, la fonction retenue pour le cas mono-locuteur est mise en oeuvre sur les images ultérieures.

En rajoutant en prétraitement une classification de pixels permettant d'initialiser le modèle d'apparence statique en plus du modèle local de commissures des lèvres, l'approche retenue donne des résultats satisfaisants dans le cas multi-locuteurs (figure 5.1).

Figure 5.1 :
Exemples de segmentation de lèvres par notre méthode.

Enfin, la qualité d'une segmentation est généralement évaluée par des critères quantitatifs mesurant l'écart entre les points de contrôle détectés et leurs positions réelles sur des images où celles-ci sont connues (car annotées manuellement). Cette évaluation a été menée mais, si elle est objective, elle ne dit néanmoins pas si la segmentation a été capable de restaurer les mouvements labiaux de la parole avec réalisme.

Nous avons donc procédé à une évaluation subjective en testant l'intelligibilité d'un clone synthétisé à partir de l'analyse dans un cas mono-locuteur. Pour cela, l'apport pour la compréhension de numéro de téléphone de notre schéma d'analyse/synthèse a été mesuré, ce qui a mené à la conclusion que notre schéma d'analyse/synthèse était suffisamment précis pour apporter une information pertinente.

Les principales originalités de ce travail sont :

- la séparation faite entre les variations d'apparence induites par l'identité et le contexte et celles liées au mouvement et à la parole.
- l'association non-linéaire entre la forme et l'apparence.
- l'évaluation subjective.

5.2 Perspectives

Si notre méthode donne des résultats robustes et précis sur la plupart des images sur lesquelles elle a été testée, diverses améliorations pourraient être apportées pour élargir encore son champ d'application. Les améliorations possibles sont de plusieurs natures : des développements à court et moyen terme, et d'autres à plus long terme.

5.2.1 Pistes à court terme

Tout d'abord, notre évaluation qualitative a été menée dans un cadre mono-locuteur. Il serait intéressant de faire l'acquisition de séquence vidéo de numéros de téléphone avec différentes personnes pour valider totalement notre schéma d'analyse/synthèse.

Ensuite, nos algorithmes ont été développés sous environnement Matlab ce qui conduit à des temps de calcul relativement important. Une programmation intégrale en C++, permettrait de se rapprocher des impératifs pour un traitement en temps réel, pré-requis indispensable pour de nombreuses applications.

Par ailleurs, si les problèmes d'éclairage ont pris en compte dans notre travail, notre méthode éprouverait des difficultés à converger dans des conditions très distinctes de celles présentes dans l'apprentissage. Outre la possibilité d'augmenter la taille de la base de donnée, le recours au filtrage rétine permettrait de gérer en partie ces problèmes dans la partie analyse.

Enfin, pour citer une piste plus à moyen terme, nous nous sommes placé pour ce travail dans un cadre 2D en vue de face. Quelques travaux ont été mené au LIS sur la détermination de l'orientation du visage ([Stillittano, 2006]), classant un visage dans cinq classes possibles (vue de face, vues de profil et deux vues intermédiaires). En construisant un modèle 2D pour chaque orientation, on pourrait obtenir une méthode s'appliquant à beaucoup plus d'images.

5.2.2 Pistes à long terme

Parmi les principales limitations des approches par modèles actifs on trouve le temps nécessaire à obtenir les données d'apprentissage et la représentativité de ces données par rapport à la réalité. Dans notre cas, la taille de la base d'apprentissage, relativement modeste, limite les capacités d'adaptation de notre modèle, qui ne pourrait pas fonctionner sur des formes de lèvres très particulières ou sur des personnes ayant des pigmentations de peau différentes.

Pour réduire ses limitations, il faudrait une méthode permettant d'annoter les images d'apprentissage autrement que manuellement, ce qui permettrait de réduire le temps consacré à cette tâche relativement pénible et d'avoir des bases de très grandes tailles. Le paradoxe étant évidemment que si l'on est capable de segmenter automatiquement les images, c'est que l'on a *a priori* déjà résolu le problème.

Un compromis verrait l'opérateur chargé de créer la base de donnée utiliser un algorithme de segmentation des lèvres et valider ou non le résultat, déplaçant le cas échéant manuellement les points de contrôle mal placé. Pour cela, il serait possible d'avoir recours à un modèle analytique des lèvres, certains donnant des résultats satisfaisants dans des situations très variables. Néanmoins, peu de ces méthodes détectent efficacement les contours intérieurs de la bouche et il faudrait donc résoudre ce problème en premier lieu.

Enfin, dès que la base aurait déjà atteint une certaine taille, le modèle actif entraîné sur les images déjà disponibles pourrait se substituer au modèle analytique. En utilisant les techniques de décomposition en composantes singulières, le modèle pourrait prendre en compte progressivement chaque nouvelle image rajoutée à la base de données.

On pourrait aussi envisager de limiter encore plus l'intervention de l'opérateur et tendre vers l'automatisation totale en définissant un critère de fiabilité qui déterminerait si la segmentation est bonne ou non. Néanmoins, obtenir un tel critère serait une tâche ardue.

Enfin, pour continuer sur le thème des limites d'application de notre méthode (et des modèles actifs d'apparence en général), on a vu que le changement de caméra et de réglages colorimétriques posait des problèmes, même si l'on a proposé une approche permettant de le réduire quelque peu en recentrant les données. Arriver à déterminer puis à normaliser les paramètres de colorimétrie de la caméra permettrait de contourner le problème, mais cela présente également un sujet d'étude ambitieux.

ANNEXE 1

Séquences Mono-Locuteur

Le tableau A.1 liste les numéros de téléphone prononcés dans les séquences mono-locuteur avec une élocution normale. Le tableau A.2 liste, quant à lui, les séquences en élocution murmurée. Les numéros de téléphone ont été déterminés de façon à ce que tous les nombres entre 0 et 99 soient cités au moins une fois.

Série d'image	Numéro de téléphone				
hl_mcf_tel002	02	77	52	26	04
hl_mcf_tel003	05	72	92	14	75
hl_mcf_tel004	06	64	19	24	59
hl_mcf_tel005	02	80	21	55	82
hl_mcf_tel006	04	63	10	47	33
hl_mcf_tel007	01	03	70	85	24
hl_mcf_tel008	03	05	76	97	54
hl_mcf_tel009	02	11	49	45	43
hl_mcf_tel010	04	76	27	04	41
hl_mcf_tel011	05	07	71	52	98
hl_mcf_tel012	02	53	53	33	08
hl_mcf_tel013	01	67	10	63	63
hl_mcf_tel014	04	02	34	75	88
hl_mcf_tel015	04	30	42	96	28
hl_mcf_tel016	06	08	70	37	44
hl_mcf_tel017	03	42	90	25	76
hl_mcf_tel018	04	66	76	97	48
hl_mcf_tel019	05	59	27	72	24
hl_mcf_tel020	01	78	06	75	28
hl_mcf_tel023	06	84	73	65	36

Tableau A.1 :
Liste des séquences et des numéros de téléphone correspondants en élocution normale.

Série d'image	Numéro de téléphone				
hl_mcf_tel024	05	10	20	04	82
hl_mcf_tel025	04	52	67	69	91
hl_mcf_tel026	04	03	22	76	60
hl_mcf_tel027	02	56	07	40	72
hl_mcf_tel028	06	89	15	09	79
hl_mcf_tel029	05	52	91	59	49
hl_mcf_tel030	02	67	23	92	17
hl_mcf_tel031	01	73	38	92	97
hl_mcf_tel032	03	69	58	32	57
hl_mcf_tel033	02	53	12	61	28
hl_mcf_tel037	02	49	47	63	84
hl_mcf_tel038	03	81	93	73	46
hl_mcf_tel039	05	10	06	72	83
hl_mcf_tel040	06	94	76	99	27
hl_mcf_tel041	01	08	76	95	74
hl_mcf_tel042	06	50	82	24	62
hl_mcf_tel043	04	39	13	31	47
hl_mcf_tel044	04	87	03	35	29
hl_mcf_tel045	03	91	68	59	18
hl_mcf_tel047	06	53	86	59	16

Tableau A.2 :
Liste des séquences et des numéros de téléphone correspondants en élocution chuchotée.

ANNEXE 2

Algorithme de Descente du Simplex

Cette méthode d'optimisation a une interprétation géométrique qui en rend facile la compréhension. Si le problème à résoudre est fonction de M variables, le simplex sera une figure géométrique à M+1 sommets correspondant aux points où sera estimée la fonction à minimiser. Par exemple dans le cas d'une fonction à deux variables, le simplex sera un triangle dans l'espace des paramètres. Dans le cas d'une fonction à trois variables, nous aurons un tétrahédon.

Si on note $\boldsymbol{e}_0$ un point initial du Simplex choisi de façon pertinente, les autres M points $\boldsymbol{e}_i$ sont obtenus de la façon suivante:

$$\boldsymbol{e}_i = \boldsymbol{e}_0 + l_i \boldsymbol{n}_i,\ 1 \le i \le M \qquad \text{(eq. A.1)}$$

où les $\boldsymbol{n}_i$ sont des vecteurs définissant une base orthonormée dans l'espace des paramètres et les l_i sont la longueur définissant l'intervalle de recherche du minimum dans chaque direction.

Une fois le Simplex initial déterminé, la fonction est calculée en chacun de ces points et l'on repère le point qui correspond au minimum (qui est donc la première estimation des paramètres minimisant le problème), celui qui correspond au maximum et enfin celui qui correspond au deuxième maximum.

La figure A.1 présente les diverses transformations classiques d'un Simplex dans le cas d'un problème à trois dimensions.

Lors du processus d'estimation, la première transformation effectuée sur le Simplex est une réflexion : la fonction est calculée au point $\boldsymbol{t}_1$ qui est le symétrique du maximum par rapport à l'hyperplan définit par les autres points du Simplex.

Si l'évaluation en $\boldsymbol{t}_1$ donne une valeur comprise entre le minimum et le second maximum, $\boldsymbol{t}_1$ devient un nouveau point du Simplex, l'ancien second maximum devenant le nouveau maximum.

Si $\boldsymbol{t}_1$ donne une valeur améliorant le minimum, on tente alors une expansion dans la même direction et on calcule la fonction au point $\boldsymbol{t}_2$. Si $\boldsymbol{t}_2$ améliore le minimum, il le remplace.

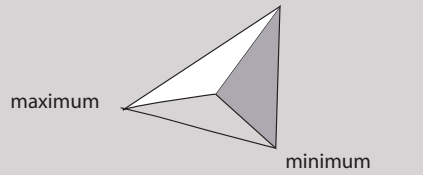
Si l'évaluation en $\boldsymbol{t}_1$ donne une valeur comprise entre le maximum et le second maximum, on effectue une contraction dans la même direction et on calcule la fonction en $\boldsymbol{t}_3$. Si $\boldsymbol{t}_3$ améliore le maximum, il le remplace.

Enfin si l'évaluation en t_1 donne une valeur supérieure au maximum, on effectue une contraction du Simplex et on évalue la fonction en t_4. Si t_4 améliore le maximum, il le remplace.

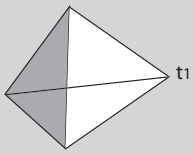
Si ni t_3 ni t_4 ne constituent une amélioration de la valeur du maximum, on procède alors à une contraction multiple, tous les points du Simplex se rapprochant du minimum et on doit alors réévaluer la fonction en tous ces points pour trouver le nouveau maximum avant de réitérer le processus.

Il est à noter que dans le cas où le problème est à grande dimension, cette opération de contraction multiple conduit à un grand nombre d'évaluations de la fonction, ce qui peut demander beaucoup de temps de calcul. De plus, cette opération est effectuée "en désespoir de cause" et on n'a aucune certitude que l'on se rapproche bel et bien du minimum, le risque de tomber dans un minimum local existant alors.

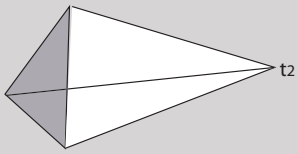
Ce problème peut-être contourné en procédant à de nouvelles opérations, proposées dans [Kaczmarczyk, 1999] qui consistent à chercher des points améliorant le maximum sur la droite joignant le maximum actuel et le minimum.



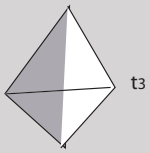
Simplex initial évaluant une fonction f de 3 variables



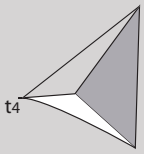
Réflexion



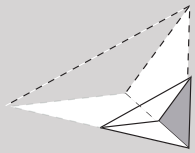
Réflexion puis expansion



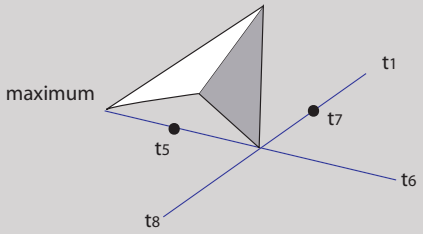
Réflexion puis contraction



Contraction



Contraction multiple



Si $f(t_1) \geq$ maximum tests aux points t_5 et t_6

Si $f(t_1) <$ maximum tests aux points t_7 et t_8

Figure A.1 :
Transformations possibles du Simplex.

ANNEXE 3

Réseaux de Neurones à Rétropropagation

Présentation rapide des réseaux de neurones

La méthodologie du réseau de neurones est initialement inspirée par les systèmes nerveux biologiques. En effet, les organismes vivants, même élémentaires, parviennent à traiter efficacement des problèmes souvent très compliqués ce qui a justifié la démarche scientifique de mimer certaines propriétés des systèmes naturels. Les possibilités des réseaux de neurones sont multiples, puisqu'ils sont en mesure d'apprendre et de s'adapter à une très grande variété de problèmes comme par exemple la reconnaissance de signal, l'identification de processus, ou dans notre cas la modélisation d'un système.

Un réseau de neurones est défini par la manière dont ses éléments sont connectés et par la force de ces connexions, ou poids. Dans le cas où l'on dispose de données d'apprentissage d'entrées et des sorties correspondantes (qui seront les cibles à atteindre pour le réseau de neurones), les poids peuvent être ajustés selon diverses règles d'apprentissage supervisé, jusqu'à ce que le réseau soit capable d'effectuer la tâche désirée (voir figure A.2).

Si l'on prend l'exemple d'un réseau de neurones à une couche avec p entrées et n sorties (figure A.3), un neurone j ($1 \leq j \leq n$) reçoit la somme des entrées E_k ($1 \leq k \leq p$) pondérée par les coefficients $w_{i,j}$ ce qui donne le potentiel P_j du neurone. Enfin la sortie O_j est égale à $f(P_j)$ avec f une fonction, fréquemment non-linéaire, telle une sigmoïde.

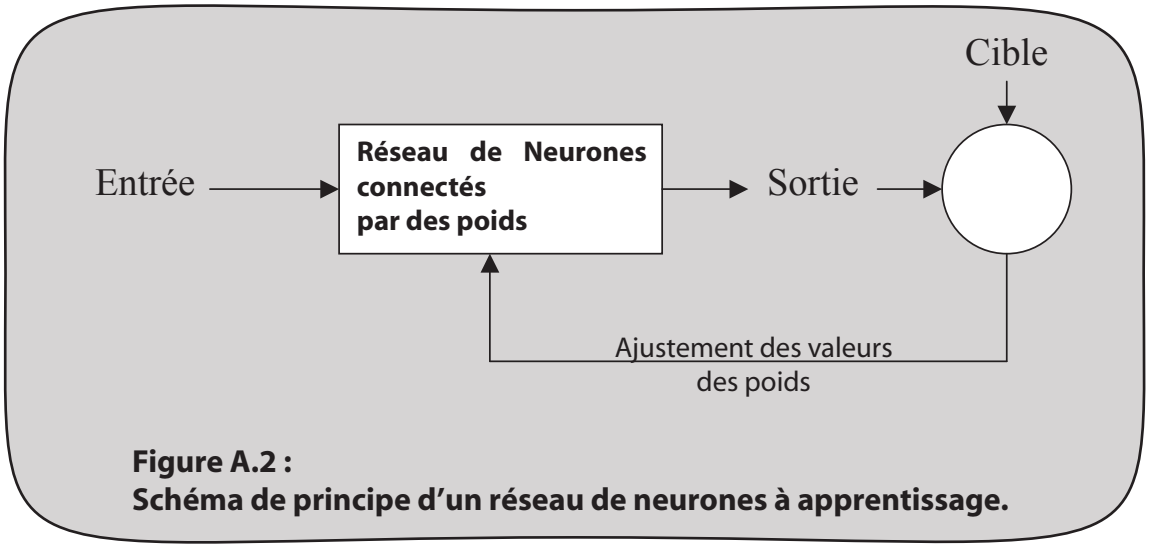


Figure A.2 :
Schéma de principe d'un réseau de neurones à apprentissage.

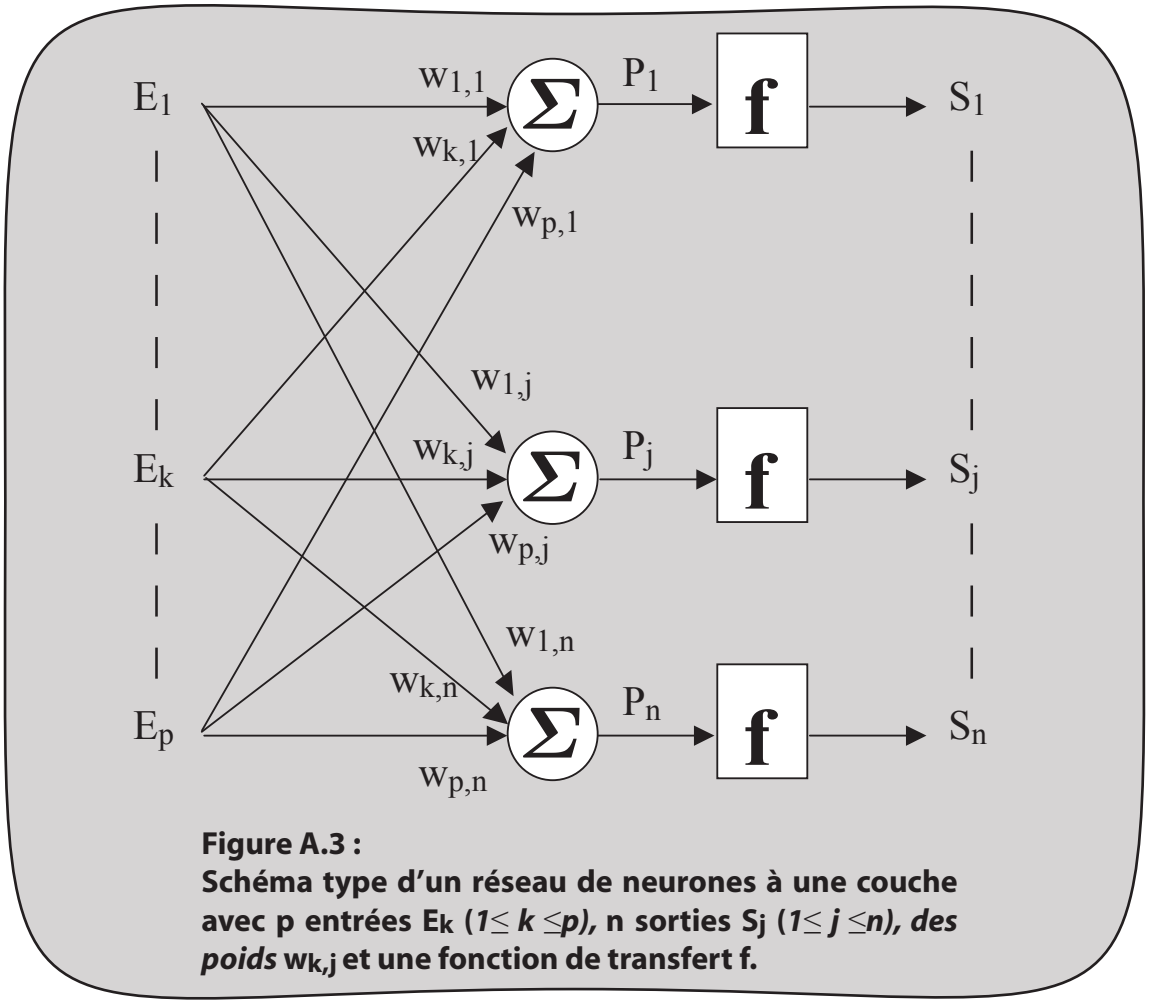


Figure A.3 :
Schéma type d'un réseau de neurones à une couche avec p entrées E_k (*1≤ k ≤p),* n sorties S_j (*1≤ j ≤n), des poids* $w_{k,j}$ et une fonction de transfert f.

Si l'on nomme C_j la sortie désirée (la cible) d'un neurone, on peut quantifier la qualité de représentation du réseau de neurones en calculant par exemple l'erreur quadratique moyenne entre les sorties attendues et celles obtenues :

$$Err = \frac{1}{2}\sum_{j=1}^{n}\left(C_j - S_j\right)^2 = \frac{1}{2}\sum_{j=1}^{n}\left(C_j - f\left(P_j\right)\right)^2 = \frac{1}{2}\sum_{j=1}^{n}\left(C_j - f\left(\sum_{k=1}^{p} w_{k,j}E_k\right)\right)^2 \quad \text{(eq. A.2)}$$

Cette erreur étant fonction des poids $w_{i,j}$ on peut la différencier et en déduire une mise à jour pour chacun des poids grâce au principe d'optimisation de descente du gradient. C'est la règle de Widrow-Hoff :

$$\Delta W_{k,j} = -g\left(C_j - S_j\right)f'\left(P_j\right)E_k \quad \text{(eq. A.3)}$$

où g est un gain d'adaptation positif à fixer.

Principe de la rétropropagation ou feed forward backpropagation

Si les réseaux à une couche de neurones peuvent correctement traiter des problèmes linéaires, rajouter des couches permet à un réseau de résoudre des problèmes non-linéaires.

Un tel réseau est donc capable d'apprendre, par l'exemple, à identifier une relation non-linéaire entre des ensembles connus d'entrées et de sorties, établissant ainsi un modèle quasi-paramétrique non-linéaire du système observé.

Ces réseaux sont construits grâce à des algorithmes de rétro-propagation du gradient (feed forward backpropagation) qui sont une extension de la règle de Widrow-Hoff. Le nom "rétropagation" vient du fait que l'observation de l'erreur sur les neurones se propage de la sortie vers l'entrée ([Hérault, 1994]).

Le principe consiste donc à remonter couche par couche à partir de l'erreur observée sur les neurones de sortie vers les neurones d'entrées et de modifier les poids synaptiques en amont de chaque couche, de manière à diminuer l'erreur commise en sortie. Le processus est itératif : à chaque itération, l'erreur globale diminue.

La méthode classique d'optimisation des poids du réseau est celle de la descente du gradient (qui consiste à suivre la ligne de plus grande pente de la surface d'erreur), néanmoins elle est de plus en plus rarement utilisée. En effet, en plus de nécessiter un très grand nombre d'itérations, elle est susceptible de tomber dans des minima locaux (ces minima étant dus au caractère non-linéaire) et donc de ne pas converger vers un modèle pleinement satisfaisant et nécessite en outre de déterminer un gain d'adaptation adéquat.

La méthode la plus communément utilisée actuellement est celle de Levenberg-Marquardt ([Hagan, 1994]). La différence fondamentale par rapport à la méthode de descente du gradient classique est que les dérivées secondes sont prises en compte. En effet, si la dérivée (ou le gradient) peut donner la direction à suivre pour améliorer les poids, la dérivée seconde (liée au rayon de courbure de la surface des erreurs) permet également de connaître le pas (ce qui correspond au gain d'adaptation de la descente du gradient) et donc de converger de façon quadratique vers la solution.

REFERENCES BIBLIOGRAPHIQUES

[Abboud, 2004] B. Abboud, F. Davoine,"Appearance Factorization Based Facial Expression Recognition and Synthesis", *17th International Conference on Pattern Recognition (ICPR'04)*, pp. 163-166, Cambridge, Royaume-Uni, 2004.

[Aleksic, 2005] P.S. Aleksic et A.K. Katsaggelos, "Automatic Facial Expression Recognition Using Facial Animation Parameters and Multi-Stream HMMs", *6th international Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05)*, Montreux, Suisse, 2005.

[Arulampalam, 2002] S. Arulampalam, S.R. Maskell, N.J. Gordon et T. Clapp, "A Tutorial on Particle Filters for On-Line Nonlinear/non-Gaussian Bayesian Tracking", *IEEE Trans. On Signal Processing*, Volume 50, no. 2, pp. 174-188, 2002.

[Ayoujil, 2003] H. Ayoujil, "Synthèse d'Avatar par Modèle d'Apparence et Triangulation", Rapport de Stage Ingénieur (ENSERG, INPG), Grenoble, 2003.

[Bailly, 2001] G. Bailly, "Audiovisual Speech Synthesis", *In ETRW on Speech Synthesis*, P. Taylor Editor, Pershire, Ecosse, 2001.

[Bailly, 2003] G. Bailly, F. Elisei, M. Odisio, D. Pelé, D. Caillièr et K. Grein-Cochard, "Talking faces for MPEG-4 compliant scalable face-to-face telecommunication", *Proceedings of the Smart Objects Conference*, pp. 204-207, Grenoble, France, Mai 2003.

[Baker, 2001] S.Baker et I.Matthews, "Equivalence and Efficiency of Image Alignment Algorithms", *In Computer Vision and Pattern Recognition Conference 2001*, Volume 1, pp. 1090–1097, 2001.

[Beaudot, 1994] W.H.A. Beaudot, "The Neural Information Processing in the Vertebrate Retina: A Melting Pot of Ideas for Artificial Vision", Thèse de Doctorat, INPG (France) Décembre 1994.

[Benoit, 2005] A. Benoit , A. Caplier, "Motion Estimator Inspired From Biological Model

For Head Motion Interpretation", *European Workshop on Image Analysis for Interactive Multimedia Services (WIAMIS 2005)*, Montreux, Suisse, Avril 2005.

[Blender] Logiciel Open Source disponible sur http://www.blender3d.org/.

[Carbini, 2003] S. Carbini, "Evaluation Quantitative d'Algorithmes de Détection Labiale : Méthodologie et Résultats", Rapport de Projet de DEA SIPT, Grenoble, 2003.

[Carson, 2002] C. Carson, S. Belongie, H. Greenspan, J. Malik, "Blobworld : Image Segmentation Using Expectation-Maximization and its Application to Image Querying", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 24, No. 8, pp. 1026-1038, Août 2002.

[Cootes, 1992] T. F. Cootes, C. J.Taylor, D. Cooper, et J. Graham, "Training Models of Shape from Sets of Examples", *In D. Hogg and R. Boyle, editors, 3rd British Machine Vision Conference*, pp. 9–18. Springer-Verlag, Septembre 1992.

[Cootes, 1993] T. F. Cootes et C. J. Taylor, "Active Shape Model Search Using Local Grey-Level Models: A Quantitative Evaluation", *In J. Illingworth, editor, 4th British Machine Vision Conference*, pp. 639–648, BMVA Press, Septembre 1993.

[Cootes, 1995] T.F. Cootes, C.J. Taylor, et D.H. Cooper, "Active Shape Models - Their Training and Application", *Computer Vision and Image Understanding*, Volume 61, No. 1, Janvier, pp. 38-59, 1995.

[Cootes, 1997] T. F. Cootes et C. J. Taylor, "A Mixture Model for Representing Shape Variation", *In A. Clarke, editor, 8th British Machine Vision Conference*, pp 110–119. BMVA Press, Essex, Septembre 1997.

[Cootes, 1998] T. F. Cootes, G.J. Edwards, et C.J. Taylor. "Active Appearance Model", *Proc. European Conference on Computer Vision 1998 (H. Burkhardt and B. Neumann Ed.s)*, Volume 2, pp. 484-498, Springer, 1998.

[Cootes, 1998+] T. F. Cootes, G. J. Edwards et C. J. Taylor, "A Comparative Evaluation of Active Appearance Model Algorithms", *In P. Lewis and M. Nixon, editors, 9th British Machine Vision Conference*, Volume 2, pp. 680–689, Southampton, Royaume-Uni, Septembre 1998.

[Cootes, 2000] T. F. Cootes, K. N. Walker et C. J. Taylor, "View-Based Active Appearance Models", *In 4th International Conference on Automatic Face and Gesture Recognition 2000*, pp. 227–232, Grenoble, France, 2000.

[Cootes, 2001] T. F. Cootes et C. J. Taylor, "Constrained Active Appearance Models", *In*

8th International Conference on Computer Vision, Volume 1, pp. 748–754, IEEE Computer Society Press, Juillet 2001.

[Cootes, 2002] T.F. Cootes et P. Kittipanya-Ngam, "Comparing Variations on the Active Appearance Model Algorithm", *In British Machine Vision Conference*, Cardiff University, pp. 837-846, Septembre 2002.

[Cootes, 2004] T. F. Cootes. "Statistical Models of Appearance for Computer Vision", Rapport Technique en Ligne Disponible sur http://www.isbe.man.ac.uk/bim/refs.html, 2004.

[Cosker, 2004] D.P. Cosker, A.D. Marshall, P.L. Rosin et Y.A. Hicks, "Speech-Driven Facial Animation using a Hierarchical Model", *IEE Proc.-Vis. Image Signal Processing*, Volume 151, No. 4, Août 2004.

[Daubias, 2002] P. Daubias, "Modèles A Posteriori de la Forme et de l'Apparence des Lèvres pour la Reconnaissance Automatique de la Parole Audiovisuelle", Thèse de Doctorat, Université du Maine, 2002.

[De Otalora, 2005] R. M. B. De Otalora, S. Herrmann, P. Zuber et W. Stechele, "An Efficient Approach for Fine-Tuning and Tracking of Face Objects", *European Workshop on Image Analysis for Interactive Multimedia Services (WIAMIS 2005)*, Montreux, Suisse, Avril 2005.

[Daugman, 1980] J.G. Daugman, "Two-Dimensional Spectral Analysis of Cortical Receptive Field Profiles", *Vision Research*, Volume 20, pp. 847–856, 1980.

[Delmas, 2000] P. Delmas, "Extraction des Contours de Lèvres d'un Visage Parlant par Contours Actifs, Application à la Communication Multimodale", Thèse de Doctorat, INPG (France), 2000.

[Delmas, 2002] P. Delmas, N.Eveno, et M. Lievin, "Towards Robust Lip Tracking", *International Conference on Pattern Recognition (ICPR'02)*,Québec City, Canada, Août 2002.

[Dodd, 1987] B. Dodd et R. Campbell. "Hearing by Eye: The Psychology of Lipreading". *Lawrence Erlbaum Associates*, Londres, 1987.

[Dornaika, 2004] F. Dornaika, F. Davoine et M. Dang, "A Stochastic Approach for Appearance-Based 3D Face Tracking", *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, April.

[Dupont, 2000] S. Dupont et J. Luettin, "Audio-Visual Speech Modeling for Continuous

Speech Recognition", *IEEE Transactions on Multimedia*, Volume 2, No. 3, septembre 2000.

[Duta, 1997] N. Duta et M. Sonka, "An Improved Active Shape Model: Handling Occlusion and Outliers", *In 9th International Conference on Image Analysis and Processing*, pp. 398-405, Florence, Italie, Septembre 1997.

[Ekman, 1978] P. Ekman et W.V Friesen, "Facial Action Coding System", *Consulting Psychologists Press Inc.*, 577 College Avenue, Palo Alto, California 94306, 1978.

[Eveno, 2003] N. Eveno, "Repérage d'Indices Visuels Pertinents dans le Visage", Thèse de Doctorat, INPG (France), 2003.

[Eveno, 2004] N.Eveno, A. Caplier, et P-Y Coulon, "Automatic and Accurate Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 14, No.5, pp. 706-715, Mai 2004.

[Ezzat, 2002] T. Ezzat, G. Geiger and T. Poggio, "Trainable Videorealistic Speech Animation," *ACM Transactions on Graphics*, Volume 21, No. 3, pp. 388-398, 2002.

[Faruquie, 2000] T. A. Faruquie, A. Majumdar, N. Rajput et L. V. Subramaniam, "Large Vocabulary Audio-Visual Speech Recognition using Active Shape Models", *In Proc. International Conference on Pattern Recognition (ICPR 2000)*, Barcelone, Espagne, Septembre 2000.

[Ford, 1998] A. Ford et A. Roberts, "Color Space Conversion", Rapport Technique Disponible en Ligne sur http://inforamep.net/poyton/PDFs/coloureq.pdf.

[Gacon, 2004] P. Gacon, P.-Y. Coulon et G. Bailly, "Shape and Sampled-Appearance model for Mouth Components Segmentation", *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'04)*, Lisbonne, Portugal, 2004.

[Gacon, 2005] P. Gacon, P.-Y. Coulon et G. Bailly. "Statistical Active Model for Mouth Components Segmentation", *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, Philadelphie, USA, 2005.

[Gacon, 2006] P. Gacon, P.-Y. Coulon et G. Bailly. "Audiovisual Speech Enhancement Experiments for Mouth Segmentation Evaluation", *2006 14th European Signal Processing Conference (EUSIPCO'06)*, Florence, Italie, 2006.

[Gagné, 1997] J.-P. Gagné et L. Boutin, "The Effects of Speaking Rate on Visual Speech Intelligibility", *In European Tutorial Workshop on Audio-Visual Speech Processing,* Rhodes, Grèce, pp. 29-32, 1997.

[Gross, 2004] R. Gross, I. Matthews et S. Baker, "Constructing and Fitting Active Appearance Models With Occlusion", *In IEEE Workshop on Face Processing in Video*, Juin 2004.

[Hagan, 1994] Hagan, M. T., and M. Menhaj, "Training Feedforward Networks with the Marquardt Algorithm", *IEEE Transactions on Neural Networks*, Volume 5, No. 6, pp. 989-993, 1994.

[Hall, 2000] D. Hall, V. Colin de Verdière et J. L. Crowley, "Object Recognition using Coloured Receptive Fields", *In Proc. Of the Euopean Conference on Computer Vision*, pp. 164-177, Dublin, Irlande, 2000.

[Hammal, 2004] Z. Hammal, A. Caplier, "Eyes and Eyebrows Parametric Models for Automatic Segmentation", *In Proc. IEEE Southhwest Symposium on Image Analysis and Interpretation*, Nevada, Etats-Unis, 2004.

[Hennecke, 1994] M. Hennecke, V. Prasad, et D. Stork. "Using deformable templates to infer visual speech dynamics", *28 th Annual Asimolar Conference on Signals, Systems, and Computer*, Volume 2, IEEE Computer, Pacific Grove, pp. 576-582, 1994.

[Hérault, 1994] J. Hérault et C. Jutten, "Réseaux Neuraux et Traitement du Signal", Hermès, Paris, 1994.

[Horbelt and Dugelay, 1995] S. Horbelt et J. L. Dugelay, "Active Contours for Lipreading – Combining Snakes with Templates", *15th GRETSI Symposium on Signal and Image Processing*, Juan les Pins, France, 1995.

[Hou, 2001] X. Hou, S. Li, H. Zhang et Q. Cheng, "Direct Appearance Model", *In Computer Vision and Pattern Recognition Conference*, Volume 1, pp. 828-833, 2001.

[Hsu, 2002] R.-L. Hsu, M. Abdel-Mottaleb et A. K. Jain, "Face Detection in Color Images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Volume 24, No. 5, pp. 696-706, Mai 2002.

[Kaczmarczyk, 1999] G. Kaczmarczyk, "Downhill Simplex Method for Many (~20) Dimension", Rapport Technique en Ligne Disponible sur http://paula.univ.gda.pl/~dokgrk/simplex.html, 1999.

[Kass, 1987] M. Kass, A. Witkin et D. Terzopoulos, "Snakes: Active Contour Models", *International Journal of Computer Vision*, pp. 321-331, 1987.

[La Cascia, 1998] M. La Cascia, J. Isidoro et S. Sclaroff, "Head tracking via Robust

Registration in Textures Map Images", *In IEEE Conf. on Computer Visions and Pattern Recognition*, Santa Barbara, CA, 1998.

[Lallouache, 1991] T. Lallouache, "Un Poste Visage-Parole. Acquisition et Traitement Automatique des Contours des Lèvres", Thèse de doctorat, INPG (France), 1991.

[Lanitis, 1994] A. Lanitis, C. Taylor et T. Cootes, "Automatic Tracking, Coding and Reconstruction of Human Faces Using Flexible Appearance Models", *IEEE Electronic Letters*, Volume 30, pp. 1578–1579, 1994.

[Le Goff, 1995] B. Le Goff, T. Guiard-Marigny et C. Benoît. "Read my Lips… and my Jaws! How Intelligible are the Components of a Speaker's Face", *In Proc. Of the European Conf. On Speech Communication and Technology*, pp. 291-294, Madrid, Espagne, 1995.

[Lievin, 2000] M. Lievin, "Analyse Entropico-Logarithmique de Séquences Vidéo Couleur, Application à la Segmentation et au Suivi de Visages Parlants", Thèse de Doctorat en Science de l'Ingénieur, INPG (France), 2000.

[Liévin, 2004] M. Liévin, F. Luthon, "Nonlinear Color Space and Spatiotemporal MRF for Hierarchical Segmentation of Face Features in Video", *IEEE Transactions on Image Processing*, Volume 13, No. 1, pp. 66-71, 2004.

[Liew, 1999] W.C. A. Liew, K.L. Sum, S.H. Leung et W.H. Lau, "Fuzzy segmentation of lip image using cluster analysis", *1999 European Conference on Speech Communication and Technology (EUROSPEECH'99)*, Hongrie, 1999.

[Liew, 2000] A.W.C. Liew, S.H. Leung et W.H. Lau, "Lip Contour Extraction Using a Deformable Model", *Int. Conf. on Image Processing (ICIP'00)*, Vancouver, Canada, 2000.

[Lindeberg, 1998] T. Lindeberg, "Feature Detection with Automatic Scale Detection", *IJVC*, Volume 30, No.2, pp. 77-116, 1998.

[Lopez Medina, 2006] A. Lopez Medina, "Interface Graphique Utilisateur sur Systèmes de Détection Automatique des Lèvres d'un Visage et sur l'Evaluation de Prétraitement Couleur pour la Détection des Lèvres d'un Visage", Rapport de Projet de Fin d'Etude, Université de Vic (Espagne) Grenoble, 2006.

[Lucas, 1981] B. Lucas et T. Kanade, "An Iterative Image Registration Technique with an Application in Stereo Vision", *In The 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[Luettin, 1996] J. Luettin, N.A. Thacker, et S.W. Beet, "Locating and Tracking Facial

Speech Features", *Proceedings of the International Conference on Pattern Recognition*, Vienne, Autriche, 1996.

[Lyons, 2003] M. J. Lyons, M. Haehnel et N. Tetsutani, "Designing, Playing, and Performing with a Vision-Based Mouth Interface", *2003 Conference on New Interfaces for Musical Expression (NIME-03)*, pp. 116-121, Montréal, Canada, 2003.

[MacLeod, 1987] A. MacLeod et Q. Summerfield, "Quantifying the Contribution of Vision to Speech Perception in Noise", *British Journal of Audiology,* Volume 21, pp. 131-141, 1987.

[Matthews, 2002] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, "Extraction of Visual Features for Lipreading", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 24, No. 2, Février 2002.

[Marr, 1980] D. Marr et E. Hildreth, "Theory of Edge Detection", *Proc. of the Royal Society of London : Biological Science*, Volume 207, pp. 187-217, 1980.

[McGurk, 1976] H. McGurk et J. McDonald, "Hearing Lips and Seeing Voices", *Nature*, pp. 746-748, décembre 1976.

[MPEG, 1997] MPEG-N1902, "Text for CD 14496-2 Video", ISO/IEC JTC1/SC29/WG11 N1886, MPEG97, Novembre 1997.

[Nascimento, 2005] J. Nascimento et J. Salvador Marques, "Adaptive Snakes Using the EM Algorithm", *IEEE Transactions on Image Processing*, Volume 14, Issue 11, pp. 1678-1686, 2005.

[Nefian, 2002] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, et K. Murphy, "A Coupled HMM for Audio-Visual Speech Recognition", *2002 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP'02)*, Volume 2, pp. 2013-2016, Orlando, Etats-Unis, Mai 2002.

[Nelder, 1965] J.A. Nelder, R. Mead, *Computer Journal*, Volume 7, pp. 308-313, 1965.

[Odisio, 2003] M. Odisio et G. Bailly, "Shape and appearance models of talking faces for model-based tracking", *Auditory-Visual Speech Processing Workshop (AVSP'03)*, France, 2003.

[Odisio, 2005] M. Odisio, "Estimation des Mouvements du Visage d'un Locuteur dans une Séquence Audiovisuelle", Thèse de Doctorat, INPG (France), 2005.

[Ostermann, 2004] J. Ostermann, A. Weissenfeld, "Face Animation for Human Computer

Interfaces", *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'04)*, Lisbon, Portugal, 2004.

[Pantic, 2000] M. Pantic et L. J.M. Rothkrtanz, "Automatic Analysis of Facial Expressions : The State of Art", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22, No. 12, Décembre 2000.

[Pantic, 2000+] M. Pantic et L. J.M. Rothkrtanz, "Expert System for Automatic Analysis of Facial Expressions", *Image and Vision Computing*, Volume 18, pp. 881-905, 2000.

[Pantic, 2006] M. Pantic et I. Patras, "Dynamic of Facial Expression : Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences", *IEEE Transactions on Systems, Man and Cybernetics*, Volume 36, No. 2, Avril 2006.

[Pandzic , 1999] I. Pandzic, J. Ostermann, et D. Millen, "Users Evaluation: Synthetic Talking Faces for Interactive Services", *The Visual Computer*, Volume 15, pp. 330-340, 1999.

[Patterson, 2002] E.K.Patterson, S. Gurbuz, Z. Tufekci et J.H. Gowdy, "Moving-Talker, Speaker-Independent Feature Study and Baseline Results Using the CUAVE Multimodal Speech Corpus", *EURASIP Journal on Applied Signal Processing*, Décembre, 2002.

[Petajan, 1984] E. Petajan, "Automatic Lipreading to Enhance Speech Recognition". Thèse de Doctorat, University of Illinois at Urbana-Champaign, 1984.

[Poggio, 1998] T. Poggio, et A. Hulbert, "Synthesizing a Color Algorithm From Examples", Science, Volume 239, pp. 482-485, 1998.

[Potamianos, 2004] G. Potamianos, C. Neti, J. Luettin et I. Matthews, "Audiovisual Automatic Speech Recognition: an Overview", *Audiovisual Speech Processing*, E. Bateson, G. Bailly and P. Perrier editors, MIT Press, 2004.

[Rabiner, 1993] L.R. Rabiner et B.H. Juang. "Fundamentals of Speech Recognition". Prentice Hall, Englewood Cliffs, NJ, 1993.

[Reisberg, 1987] D. Reisberg, J. McLean, et A. Goldfield. (1987), "Easy to Hear but Hard to Understand: a Lipreading Advantage with Intact Auditory Stimuli", *In Hearing by Eye: The Psychology of LipReading*, B. Dodd and R. Campbell, Editors, Lawrence Erlbaum Associates: Hillsdale, New Jersey, pp. 97-113, 1987.

[Revéret, 1999] L. Revéret, "Conception et Evaluation d'un Système de Suivi Automatique des Gestes Labiaux en Parole", Thèse de Doctorat, Institut National Polytechnique, Grenoble, 1999.

[Romdhani, 1999] S. Romdhani, S. Gong, et A. Psarrou, "A Multi-View Non-Linear Active Shape Model Using Kernel PCA", *In T. Pridmore and D. Elliman, editors, 10th British Machine Vision Conference*, Volume 2, pp. 483–492, Nottingham, Royaume-Uni, Septembre, 1999.

[Romdhani, 2005] S. Romdhani, "Face Image Analysis Using a Multiple Features Fitting Strategy", Thèse de Doctorat, Faculté de Basel, Belgique, 2005.

[Rowley, 1996] H.A. Rowley, S. Baluja et T. Kanade, "Neural Network-Based Face Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 20, No. 1, pp. 23-38, Janvier 1996.

[Schapire, 1997] R.E. Schapire, Y. Freund, P. Bartlett et W.S. Lee, "Boosting the margin : a new explanation for the effectiveness of voting methods", *In Proc. 14th International Conference on Machine Learning (ICML)*, pp. 322–330, 1997.

[Schwartz, 2002] J.-L. Schwartz, D. Sodoyer, L. Girin, J. Klinkisch, et C. Jutten. "Separation of Audio-Visual Speech Sources: A New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli", *EURASIP Journal on Applied Signal Processing*, 11, pp. 1165-1173, 2002.

[Schwartz, 2004] J.-L. Schwartz, F. Berthommier et C. Savariaux. "Seeing to Hear Better: Evidence for Early Audio-Visual Interactions in Speech Cdentification", *In Cognition,* Volume 93, pp. 69-78, 2004.

[Scholkopf, 1998] B. Scholkopf, A. Smola et K. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation*, Volume 10, pp. 1299-1319, Mai 1998.

[Schmid, 2000] C. Schmid, R. Mohr et C. Bauckhage, "Evaluation of Interest Point Detector", International Journal of Computer Vision, Volume 37, pp. 151-172, 2000.

[Sozou, 1994] P. Sozou, T. F. Cootes, C. J. Taylor, et E. D. Mauro, "A Non-Linear Generalisation of PDMs Using Polynomial Regression", *In E. Hancock, editor, 5th British Machine Vision Conference*, pp. 397–406, York, Angleterre, Septembre 1994.

[Sozou, 1997] P. Sozou, T. F. Cootes, C. J. Taylor, et E. D. Mauro, "Non-Linear Point Distribution Modelling Using a Multi-Layer Perceptron", *In IVC*, No. 15, pp. 457–463, 1997.

[Stillittano, 2005] S. Stillittano, "Estimation de l'Orientation du Visage", Rapport de Stage de Master SIPT, INPG, Grenoble, 2005.

[Sugimoto, 2004] A. Sugimoto, M. Kimura et T. Matsuyama, “Detecting Human Heads and Face Orientations under Dynamic Environment”, *Third Int. Workshop on Articulated Motion and Deformable Objects (AMDO’04)*, pp. 163-179, Palma de Mallorca, Espagne, 2004.

[Summerfield , 1989] A. Summerfield, A. MacLeod, M. McGrath, et M. Brooke, “Lips, Teeth, and the Benefits of Lipreading”, *In Handbook of Research on Face Processing*, A.W. Young et H.D. Ellis, Editors, Elsevier Science Publishers, Amsterdam, pp. 223-233, 1989.

[Talle, 1997] B. Talle, G. Krone, G. Palm, “Comparison of Neural Architectures for Sensorfusion”, *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’97)*, Munich, Allemagne, 1997.

[Tian, 2000] Y. Tian, T. Kanade et J. Cohn, “Robust Lip Tracking by Combining Shape, Color and Motion”, *4th Asian Conference on Computer Vision (ACCV’00)*, Janvier, 2000.

[Vitkovitch , 1994] M. Vitkovitch et P. Barber, “Effect of Video Frame Rate on Shadowing”, *Journal of Speech and Hearing Research*, Volume 37, pp. 1204-1210, 1994.

[Vogelhuber, 2000] V. Vogelhuber et C. Schmid, “Face Detection Based on Generic Local Descriptors and Spatial Constraints”, *In Proc. of the International Conference on Pattren Recognition*, pp. 1084-1087, Barcelone, Espagne, 2000.

[Yang, 1998] M.-H. Yang et N. Ahuja, “Detecting Human Face in Color Images”, *International Conference on Image Processing (ICIP’98)*, Volume 1, pp. 127-130, Chicago, Octobre 1998.

[Yehia, 1998] H. Yehia, P. Rubin, et E. Vatikiotis-Bateson, “Quantitative Association of Vocal-Tract and Facial Behavior”, *Speech Communication*, Volume 26, No.1, pp. 23-43, 1998.

[Young, 1985] R.A. Young, “The Gaussian Derivative Theory of Spatial Vision : Analysis of Cortical Cell Recptive Field Line-Weighting Profiles”, *Rapport Technique GMR-4920, General Motors Resserch Laboratories*, Mai 1985.

[Young, 1995] I.T. Young, L.J. Van Vliet, “Recursive Implementation of the Gaussian Filter”, *Signal Processing,* Volume 44, pp. 139-151, 1995.

[Zhang, 2000] X. Zhang et R.M. Mersereau, “Lip Feature Extraction Towards an Automatic Speechreading System”, *In Proc. International Conference on Image Processing (ICIP’00),*

Vancouver, Canada, 2000.

[Zhang, 2002] X. Zhang, R. M. Mersereau, M. A. Clements et C. C. Broun, “Visual Speech Feature Extraction for Improved Speech Recognition”, *In Proc. ICASSP’02*, pp. 1993-1996, 2002.

[Zhang, 2003] B. Zhang, W. Gao, S. Shan et W. Wang, “Constraint Shape Model Using Edge Constraint and Gabor Wavelet Based Search”, *Audio-and Video-Based Biometrie Person Authentication, 4th International Conference, (AVBPA 2003)*, Guildford, Royaume-Uni, Juin, 2003.

[Zhang, 2005] Y. Zhang et Q. Ji, “Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 27, No. 5, Mai 2000.

PUBLICATIONS

[Gacon, 2004] P. Gacon, P.-Y. Coulon et G. Bailly, “Shape and Sampled-Appearance model for Mouth Components Segmentation”, *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS’04)*, Lisbonne, Portugal, 2004.

[Gacon, 2005] P. Gacon, P.-Y. Coulon et G. Bailly. “Statistical Active Model for Mouth Components Segmentation”, *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’05)*, Philadelphie, USA, 2005.

[Gacon, 2005] P. Gacon, P.-Y. Coulon, G. Bailly. “Non-Linear Active Model for Mouth Inner and Outer Contours Detection”, *2005 13th European Signal Processing Conference (EUSIPCO’05)*, Antalya, Turquie, 2005.

[Gacon, 2005] P. Gacon, P.-Y. Coulon, G. Bailly. “Modèle Statistique et Description Local d’Apparence pour la Détection du Contour des Lèvres”, *20e Colloque sur le traitement du signal et des images (GRETSI’05)*, Louvain-la-Neuve, Belgique, 2005.

[Gacon, 2006] P. Gacon, P.-Y. Coulon et G. Bailly. “Audiovisual Speech Enhancement Experiments for Mouth Segmentation Evaluation”, *2006 14th European Signal Processing Conference (EUSIPCO’06)*, Florence, Italie, 2006.

La segmentation de la bouche est un problème important qui trouve des applications dans plusieurs domaines du multimédia. Dans ce travail, notre objectif est d'obtenir une détection robuste et efficace des contours des lèvres de façon à être capable de restaurer les mouvements de la parole aussi fidèlement que possible. Nous apportons une attention particulière au contour intérieur de la bouche dans la segmentation est une tâche difficile à cause des variations non-linéaires de l'apparence. Nous proposons une méthode basée sur un modèle statistique de la forme et de l'apparence échantillonnée faisant intervenir des descripteurs gaussiens locaux d'apparence. Notre hypothèse est que la réponse de ces descripteurs locaux peut être prédite à partir de la forme par le biais d'un réseau de neurones non-linéaire. Nous avons d'abord testé cette hypothèse dans un cas mono-locuteur et l'avons ensuite généralisé à un cas multi-locuteurs en tenant de la variabilité inter-personne. A cet effet, nous adaptons progressivement notre modèle au locuteur traité en déterminant son apparence caractéristique. A partir de notre segmentation de la bouche, nous pouvons ensuite générer un clone de la bouche de la personne dont les mouvements seront aussi proches que possible de ceux de l'originale. Finalement, nous avons évalué quantitativement puis qualitativement la pertinence de notre méthode en menant une expérience qui a quantifié l'apport effectif de compréhension de notre schéma d'analyse/synthèse dans le cas de numéros de téléphone en milieu bruité.

Image Analysis and Shape Models for Detection and Recognition. Application to Face in Multimedia.

Mouth segmentation is an important issue which applies in many multimedia applications. In this work, our goal is to have a robust and efficient detection of lips contour in order to restore as faithfully as possible the speech movement. We specially focus on the detection of the inner mouth contour which is a difficult task due to the non-linear appearance variations. We propose a method based on a statistical model of shape and sampled-appearance with local appearance gaussian descriptors. Our hypothesis is that the response of the local descriptors can be predicted from the shape by a non-linear neural network. We tested this hypothesis with a single speaker task and then generalized it to take care of the inter person appearance variability in a multi-speaker task. To that purpose, we adapt progressively our model to the speaker by determining its characteristic appearance. From our automatic segmentation of the mouth, we can then generate a clone of a speaker mouth whose lips movements will be as close as possible of the original ones. Finally, we evaluate our method relevance quantitatively and next qualitatively by carrying out an experience which quantify the effective enhancement in comprehension brought by our analysis-resynthesis scheme in a telephone enquiry task.

Mots-clefs : Analyse des Lèvres et du Visage, Modèles Actifs de Forme et d'Apparence, Description Locale d'Apparence, Association Non-Linéaire Forme/Apparence, Evaluation Qualitative de Méthode.

Keywords : Face and Lips Analysis, Active Shape and Appearance Model, Local Appearance Description, Non-Linear Shape/Appearance Association, Qualitative Method Evaluation.

Laboratoire de Images et de Signaux
46, Avenue Félix Viallet
38031 Grenoble Cedex