Segmentation et suivi des contours externe et interne des lèvres pour des applications de maquillage virtuel et de labiophonie

Sébastien Stillittano

INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'Institut polytechnique de Grenoble

Spécialité : *Signal, Image, Parole, Télécoms*

préparée au laboratoire GIPSA-lab/DIS

dans le cadre de **l'Ecole Doctorale** *Electronique, Electrotechnique, Automatique & Traitement du Signal*

présentée et soutenue publiquement

par

Sébastien STILLITTANO

le 26/05/09

SEGMENTATION ET SUIVI DES CONTOURS EXTERNE ET INTERNE DES LEVRES POUR DES APPLICATIONS DE MAQUILLAGE VIRTUEL ET DE LABIOPHONIE

Sous la direction de Mme. **Alice CAPLIER**
Sous la co-direction de M. **Pierre-Yves COULON**

JURY

M. Maurice MILGRAM	Président et rapporteur
M. Renaud SEGUIER	Rapporteur
Mme. Alice CAPLIER	Directeur de thèse
M. Pierre-Yves COULON	Co-encadrant
M. Pierre ADAM	Encadrant industriel
M. Jean-Luc DUGELAY	Examinateur

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Ce travail de thèse CIFRE a été effectué au Département Images et Signal (DIS) du laboratoire GIPSA et financé par la société Vesalis. C'est avec plaisir que j'exprime ici mes remerciements à toutes les personnes qui ont, de près ou de loin, contribué à la réalisation de ce travail.

Tout d'abord, je tiens à remercier les membres de jury qui ont rapidement accepté les changements de dernières minutes nécessaires pour l'acceptation de la proposition du jury par le Collège Doctoral. Je voudrais donc adresser à Monsieur Maurice Milgram, professeur à l'Université Pierre et Marie Curie (Paris VI), ma plus profonde gratitude pour avoir accepté d'être à la fois président du jury et rapporteur. Je remercie Monsieur Renaud Seguier, professeur associé à Supelec (Campus de Rennes), d'avoir consacré un temps précieux à rapporter ce travail. Leus observations et leurs remarques pertinentes ont permis, en plus d'enrichir le contenu de ce mémoire, de fournir de nouvelles idées afin d'améliorer les algorithmes développées.
Mes remerciements vont également à Monsieur Jean-Luc Dugelay, professeur à Eurecom (Sophia-Antipolis), pour l'intérêt qu'il a bien voulu porter à cette thèse en l'examinant et en assistant à la soutenance.

Maintenant, je tiens tout particulièrement à remercier Alice Caplier, Maître de Conférence HDR au GIPSA-lab, qui a assuré la direction et le suivi de cette thèse avec une bienveillance permanente. Elle a su me guider et me conseiller tout au long de ces trois années afin de mener à bien ce travail. Pour sa disponibilité et sa confiance, je la remercie vivement.

J'exprime mes sincères remerciements à Pierre-Yves Coulon, Professeur au GIPSA-lab et co-directeur de cette thèse, pour l'attention qu'il a porté à cette thèse et les conseils qu'il m'a donné, dans la continuité du stage de Master que j'avais effectué sous sa direction.

Je tiens à adresser mes plus vifs remerciements à Monsieur Jean-Marc Robin, Président Directeur Général de Vesalis, pour m'avoir permis de faire une thèse au sein d'une entreprise innovante; le produit Makeuponline est une belle invention et je lui souhaite le succès qu'il mérite. D'une manière générale, je remercie toute l'équipe de Vesalis de Grenoble et Clermont-Ferrand dirigée par Christophe Blanc. Et en particulier ceux qui, à un moment donné, ont partagé mon bureau, Benoît, Jing, Pierre, Tony, Vincent.

J'associe à mes remerciements tous les membres du GIPSA-lab avec qui j'ai partagé avec plaisir ces années de recherche, notamment les doctorants de l'association des thésards. Une mention toute particulière pour Vincent Girondel qui, au cours de son post-doc effectué au DIS, m'a grandement aidé à améliorer mes résultats et à mettre de l'ordre dans mes codes.

Je remercie enfin ceux à qui je dédie cette thèse :

- *à mes parents, qui m'ont offert l'opportunité de faire les études que je souhaitais et qui ont toujours été présents pour moi,*
- *à mon épouse, pour son soutien pendant ces trois années et aussi pour m'avoir supporter pendant mes nuits blanches de rédaction. Merci de partager mon quotidien.*

Sébastien

INTRODUCTION

Ces dernières années, l'analyse des visages connaît un intérêt grandissant dans le domaine de la vision par ordinateur. Le visage est un vecteur d'information puissant de la communication entre être humains : il fournit des indications pertinentes sur l'identité d'une personne, sur son état émotionnel ou sur ce qu'elle dit.

Grâce à cet engouement et aux progrès continus des calculateurs, les méthodes d'analyse faciale se multiplient et deviennent de plus en plus précises. Elles font intervenir différentes opérations de traitement d'images telles que la segmentation de régions (peau, masque capillaire...), l'extraction des contours des traits permanents (yeux, sourcils, nez, bouche...) ou le suivi temporel du visage ou de ses traits permanents.

L'interprétation de ces informations bas niveau (régions et/ou contours du visage) intervient dans de nombreuses applications qui connaissent actuellement un essor important.

- Dans le domaine de la sécurité, on peut distinguer l'identification de personne ([Brunelli, 1995]; [Zhao, 2003]) et l'authentification ([Poh, 2001]; [Ross, 2004]) qui utilisent des caractéristiques physiologiques comme l'iris, la rétine, la forme du visage, des lèvres ou des oreilles (biométrie). L'identification consiste à retrouver l'identité d'une personne; le visage de la personne est comparé à l'ensemble de la base de données. L'authentification permet un contrôle d'accès; les caractéristiques de la personne demandant l'accès sont analysées pour confirmer son identité.
- L'interaction homme-machine est un thème de recherche qui utilise également l'analyse faciale afin de rendre l'interface entre l'utilisateur et l'ordinateur plus conviviale. Par exemple, l'analyse des émotions ([Essa, 1995]; [Hammal, 2007]) a pour but de faire réagir la machine en fonction de ce que ressent l'utilisateur ou l'animation de clone réaliste ([Terzopoulos, 1993]; [Dornaika, 2004]) permet un dialogue plus naturel.
- Dans le domaine de la réalité mixte ou augmentée, il est possible de modifier l'aspect d'un visage en le vieillissant ([Rowland, 1995]; [Lanitis, 2002]; [Wang, 2004a]), en lui ajoutant ou en lui enlevant des lunettes ([Jing, 2000]; [Park, 2005]) ou en le maquillant virtuellement [Kim, 2005] par exemple.

Toutes ces applications font intervenir une ou plusieurs zones spécifiques du visage. Parmi elles, la bouche est une région directement liée à la production de la parole, et donc à la communication humaine. La perception de la parole est bimodale, elle est à la fois acoustique et visuelle. Des études dans le domaine des sciences cognitives ont montré que le cerveau fusionne des informations auditives et visuelles pour nous aider à mieux percevoir la parole en environnement bruité ([Sumby, 1954]; [Neely, 1956]). Dans [Benoît, 1996], les auteurs montrent que l'information visuelle utile est essentiellement portée par les lèvres. Ceci explique qu'au milieu d'une foule bruyante, le fait de fixer son regard sur les lèvres de son interlocuteur permet de mieux comprendre ce qu'il dit. Cet aspect a été exploité en reconnaissance automatique de la parole, de telle manière que le développement d'un système de communication audio-visuel performant implique l'extraction préalable d'indices visuels et notamment les contours des lèvres ([Petajan, 1984]; [Potamianos, 2004]). En plus de la lecture labiale, les formes des lèvres sont une caractéristique visuelle très utile pour les applications déjà mentionnées : la reconnaissance de personne ([Jourlin, 1997]; [Wark, 1998]; [Brand, 2001]), la reconnaissance d'expressions du visage ([Tian, 2000a]; [Seyedarabi, 2006]; [Ratliff, 2008]) ou l'animation d'avatar ([Yin, 2002]; [Kuo, 2005]; [Beaumesnil, 2006]).

Ainsi, l'analyse labiale est un thème de recherche étudié de manière intensive. Mais malgré la multiplication des méthodes proposées, la segmentation des contours labiaux reste une tâche ardue. En effet, les lèvres sont hautement déformables et l'allure des contours varient significativement en fonction de l'ouverture et de l'étirement de la bouche. La couleur et la forme des lèvres (fines, épaisses ...) sont également variables en fonction de l'individu. Enfin, les changements d'illumination ou les possibles présences de barbe, de moustaches ou de rides à proximité de la bouche rendent la segmentation encore plus difficile.

Ce manuscrit présente les travaux effectués dans le cadre d'une thèse CIFRE impliquant un partenaire industriel : la PME Vesalis (http://www.vesalis.fr), et un laboratoire universitaire : le GIPSA-lab (http://www.gipsa-lab.inpg.fr). Dans le cadre de cette étude, nous nous intéressons à la segmentation des contours labiaux en rapport à deux applications :

- Un simulateur virtuel (Makeuponline) développé et édité par Vesalis, qui permet de maquiller un visage (application de fard à joue, fond de teint, mascara ...) en fonction de la position des traits permanents du visage. La détection des contours labiaux sert à appliquer le rouge à lèvre et le gloss.
- Un projet de téléphonie à l'usage des malentendants (Projet TELMA) réunissant plusieurs partenaires industriels et universitaires, dont le GIPSA-lab. La segmentation des lèvres dans ce cas est une étape utile à un module de débruitage du signal audio et à un module de production d'un signal de parole à partir d'informations visuelles (forme des lèvres et gestes de la main).

Depuis plusieurs années, le laboratoire GIPSA mène des travaux sur la segmentation automatique des traits permanents du visage pour des applications de type multimédia (réalité mixte, terminal téléphonique, interaction homme-machine, interprétation de gestes de communication non verbale, simulateur de conduite interactif…). Des études ont été proposées pour la détection de visage et l'extraction des contours des yeux, des sourcils et de l'arc mandibulaire, mais une attention particulière a été portée sur la segmentation des contours de la bouche. L'aspect bimodal de la parole a conduit les départements DPC (**D**épartement **P**arole et **C**ognition) et DIS (**D**épartement **I**mage et **S**ignal) du GIPSA-lab à mettre en œuvre très tôt des méthodes de segmentation des lèvres (cf. Fig. 0.1). Lallouache [Lallouache, 1991] extrait des paramètres labiaux en détectant la région des lèvres, préalablement maquillées en bleu, par un seuillage sur la chrominance (système *chroma-key*). Les contours actifs (ou snakes) [Delmas, 2000], les modèles paramétriques [Eveno, 2003], et les **M**odèles **A**ctifs de **F**orme (ASM) et d'**A**pparence (AAM) [Gacon, 2006] ont été utilisés lors de thèses précédentes pour la segmentation des contours labiaux.

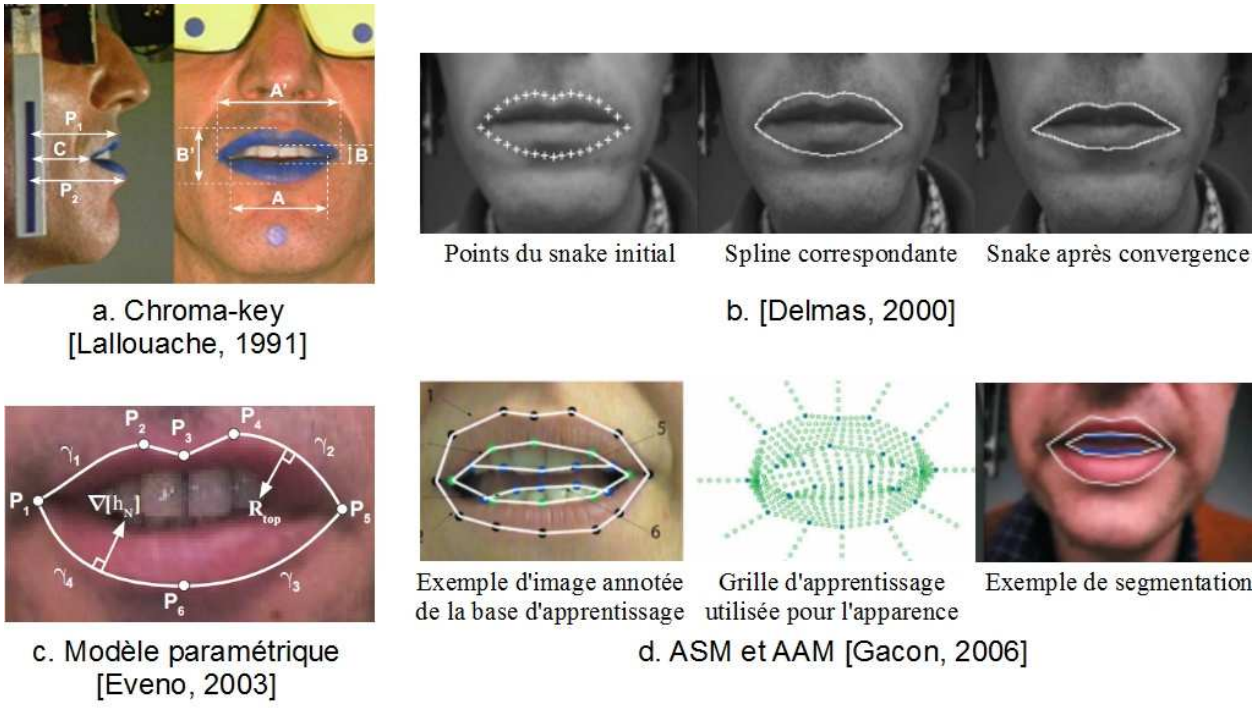


Fig. 0.1. Aperçu des méthodes de détection des contours labiaux proposées au GIPSA-lab.

En prenant en compte les deux champs applicatifs étudiés dans le cadre de cette thèse, nous avons envisagé une approche orientée contour pour extraire les mouvements des lèvres. Cette étude fait suite au travail introduit par Eveno [Eveno, 2003] sur le contour extérieur de la bouche. En plus d'améliorer la méthode proposée par Eveno en la rendant automatique et plus robuste, nous nous sommes intéressés à la détection du contour intérieur. La segmentation du contour labial interne est une tâche difficile à cause de la non linéarité du problème. L'aspect du contour interne et l'apparence de l'intérieur de la bouche peuvent varier brutalement suivant que la bouche est ouverte ou fermée, et suivant la visibilité des dents, des gencives, de la langue ou de la cavité orale (cf. Fig. 0.2).

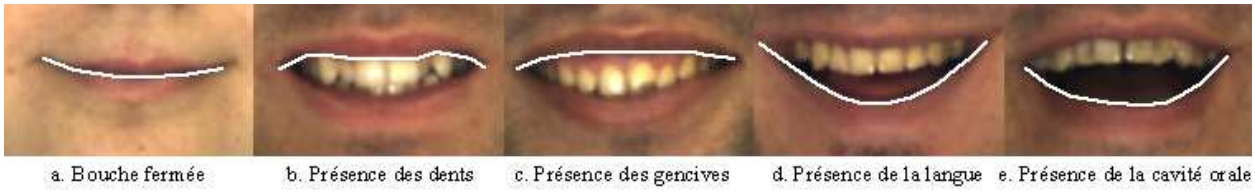


Fig. 0.2. Non linéarité de l'aspect du contour intérieur, images issues de [Martinez, 1998].
Le contour intérieur affecté par la présence d'une des caractéristiques internes est représenté en blanc.

Le chapitre 1 est consacré aux différentes applications nécessitant de connaître la forme des lèvres. Nous y détaillons notamment les deux applications directement visées par notre étude : le logiciel de maquillage virtuel et le projet TELMA. Ensuite, un éventail des nombreuses autres applications possibles est présenté en fin de chapitre.

Dans le chapitre 2, nous réalisons un état de l'art des espaces couleurs utilisés généralement pour la segmentation des lèvres et des méthodes existantes pour extraire les contours de la bouche.

- Généralement, les algorithmes de détection des contours des lèvres exploitent les espaces couleurs classiques (*RGB, YCbCr, HSV* ...), mais nous montrons qu'il existe des composantes qui permettent d'accentuer efficacement le contraste entre les lèvres et la peau.
- Le nombre important d'applications visées par les développeurs de méthodes de segmentation des lèvres a donné naissance à de nombreuses études, que nous pouvons classer en deux catégories : les techniques basées région et les techniques basées contour.

Dans le chapitre 3, nous montrons que les études bibliographiques effectuées au cours de ce travail ont permis de faire ressortir deux techniques majeures pour s'attaquer au problème de la segmentation du contour des lèvres : les contours actifs (ou snakes) et les modèles paramétriques. En les combinant, il est possible d'exploiter les propriétés des deux méthodes afin d'obtenir un algorithme robuste et rapide.

Le chapitre 4 décrit les traitements préliminaires réalisés lors de cette étude. Nous commençons par introduire les modèles paramétriques choisis pour représenter les contours extérieur et intérieur de la bouche. Ensuite, nous présentons l'algorithme du contour actif que nous avons utilisé et les différents gradients créés pour accentuer le contour labial. Finalement, un descriptif des bases d'images utilisées lors de l'évaluation de notre algorithme de segmentation est introduit à la fin de ce chapitre.

La méthode proposée pour extraire les contours des lèvres dans des images fixes est présentée dans le chapitre 5. Le déroulement de l'algorithme statique suit trois phases : la détection du visage, l'extraction de la boîte englobante de la bouche et l'extraction des contours extérieur et intérieur. Une évaluation quantitative du module statique est réalisée en comparant les résultats fournis par l'algorithme proposé avec une vérité de terrain.

Dans le chapitre 6, les quatre phases de la méthode de suivi des contours des lèvres sont décrites : le suivi de plusieurs points clefs, la réinitialisation du suivi si nécessaire, le suivi de la boîte englobante de

la bouche et l'extraction des contours extérieur et intérieur. De la même façon que pour la partie statique, une évaluation quantitative du module de suivi est réalisée en comparant les résultats avec une vérité de terrain.

Enfin, le chapitre 7 concerne l'utilisation de l'algorithme de segmentation des lèvres proposé pour nos deux applications de maquillage virtuel et de téléphonie à l'usage des malentendants. L'algorithme a également été testé lors d'une collaboration avec l'Université de Princeton (NJ, USA) dans le cadre d'une application de lecture labiale.

CHAPITRE 1

Les applications de la segmentation des lèvres

La segmentation des lèvres est un traitement qui s'intègre généralement dans un système d'interprétation plus complexe. En conséquence, une application requiert rarement de n'utiliser que les contours de la bouche, mais l'extraction des contours labiaux n'en reste pas moins une étape importante. Par exemple, les contours des lèvres couplés avec d'autres indices faciaux tels que les contours des yeux et des sourcils peuvent servir à effectuer de la reconnaissance d'émotions. De même, les systèmes de reconnaissance de la parole utilisent un traitement audio associé à un traitement visuel sur la forme des lèvres pour diminuer le taux d'erreur en environnement bruité.

En outre, les contraintes sur la forme des résultats de la segmentation ne sont pas toujours identiques selon l'application visée.
Par exemple, nous pouvons distinguer des applications qui nécessitent :

- de n'extraire que le contour extérieur des lèvres (reconnaissance d'émotions)
- de segmenter les contours extérieur et intérieur de la bouche (reconnaissance automatique de la parole)
- une segmentation très précise (identification de personne par les lèvres)
- un résultat indicatif ou seulement la position de plusieurs points clefs (animation de tête parlante)
- de détecter les contours pour des images fixes ou de suivre les contours dans des séquences vidéo.

En premier lieu, nous présentons les deux applications ayant servi de fil conducteur à ce travail de thèse : le produit Makeuponline de maquillage virtuel (cf. Section 1.1) et le projet TELMA de téléphonie à l'usage des malentendants (cf. Section 1.2). Ces deux contextes applicatifs nécessitent l'extraction de l'ensemble de la bouche (contours extérieur et intérieur) et d'obtenir une segmentation précise aussi bien pour des images fixes que pour des séquences vidéo. Ainsi, ces travaux peuvent également être utilisés par des systèmes de communication audio-visuels pour faire de la reconnaissance automatique de parole (cf. Section 1.3). Dans la section 1.4, différents thèmes de recherche, qui utilisent potentiellement les contours externes et/ou internes des lèvres, tels que la reconnaissance d'émotions, l'animation de tête parlante, l'identification de personne ou l'aide à la chirurgie plastique sont brièvement présentés.

1.1. Le projet Makeuponline™

1.1.1. Présentation de Vesalis

Ce travail de thèse, débuté en janvier 2006 et terminé en mars 2009, est un travail de thèse CIFRE financé par la société Vesalis. Vesalis est une jeune start-up française, dont le siège social se situe à Clermont-Ferrand. Le domaine de spécialisation de l'entreprise concerne principalement l'imagerie du visage pour faire de la mise en beauté virtuelle.
Les activités de Vesalis se découpent en deux secteurs distincts, le maquillage virtuel et plus récemment la biométrie faciale :

- depuis 2004, Vesalis met en œuvre une toute nouvelle approche du conseil beauté personnalisé grâce à de la réalité augmentée, à travers son produit Makeuponline™ (cf. Section 1.1.2),
- récemment, l'entreprise s'est orientée vers les marchés de la Défense et de la Sécurité Publique à travers un important projet national de biométrie faciale intitulé Biorafale, regroupant plusieurs partenaires scientifiques, dont le GIPSA-lab.

1.1.2. Le logiciel Makeuponline™

Vesalis est le développeur et l'éditeur du premier simulateur virtuel de mise en beauté automatique, temps réel et interactif (produit Makeuponline retail et .com). Makeuponline est le produit phare de la société destiné à la commercialisation. Le traitement effectué par le logiciel met en œuvre, de manière séquentielle, les phases suivantes : acquisition, détection des contours et application du maquillage.

1.1.2.a. L'acquisition et la détection des contours

Makeuponline est disponible sous deux versions distinctes, mais reposant sur la même technologie de traitement d'images :

- une version borne destinée à être placée en magasin (cf. Fig. 1.01.a)
- une version internet accessible depuis un PC ou un mobile (cf. Fig. 1.01.b).

a. Version borne b. Version internet

Fig. 1.01. Les deux versions du produit Makeuponline.

A partir des images acquises, les traits à maquiller sont extraits automatiquement. Vesalis a défini une collaboration avec le laboratoire Gipsa-lab pour cette phase de segmentation de contours. Le Gipsa-lab a développé les différentes méthodes de détection des contours des traits permanents du visage et Vesalis a effectué l'intégration et le test à grande échelle des algorithmes de détection.
La figure 1.02.a montre l'extraction des contours du visage, des yeux, des sourcils et de la bouche. Concernant les lèvres, dans la version actuelle du logiciel, seuls les contours extérieurs sont détectés, et l'algorithme de segmentation est basé sur la méthode introduite dans [Eveno, 2003] et présentée dans le chapitre 5.

a. Exemples de détection des traits permanents du visage.

b. Exemples de personnalisation.

Fig. 1.02. Détection des caractéristiques du visage.

Pour la formule borne, un même système d'acquisition et d'éclairage permet d'obtenir des conditions uniformes d'illumination pour les images acquises. Pour la solution internet, l'acquisition des images est non-contrôlée et choisie par l'utilisateur (webcam, appareil photo numérique, téléphone portable...). Les algorithmes de segmentation doivent donc être suffisamment robustes vis-à-vis des conditions d'acquisition (matériels, éclairage, pose...), mais aussi vis-à-vis de la grande diversité des personnes (forme des contours, couleur de la peau...).
Aussi, depuis le départ, il est évident que pour une très large diffusion de ce service, il faut un traitement pouvant être effectué par un PC standard, fournissant des résultats précis et fonctionnant rapidement (coût de calcul peu élevé).

1.1.2.b. L'application du maquillage

Les ingénieurs R&D et les infographistes de Vesalis ont mis en œuvre la méthode d'application de maquillage à partir des contours des traits permanents du visage obtenus lors de la phase de détection.

L'application du maquillage peut être personnalisée en fonction de la couleur des yeux et de la forme du visage (rond, carré, ovale, allongé...) détectées automatiquement (cf. Fig. 1.02.b).

Les contours du visage permettent d'appliquer du fard à joue, du fond de teint et de l'anticerne. Il est possible d'ajouter du crayon pour les yeux, du fard à paupières et du mascara. Enfin, les contours des lèvres servent à appliquer du rouge à lèvre et du gloss. Il est également possible de changer la couleur des yeux en simulant l'ajout de lentilles de contact. La figure 1.03 montre des exemples d'images obtenues après application d'un maquillage réalisé par le logiciel Makeuponline.

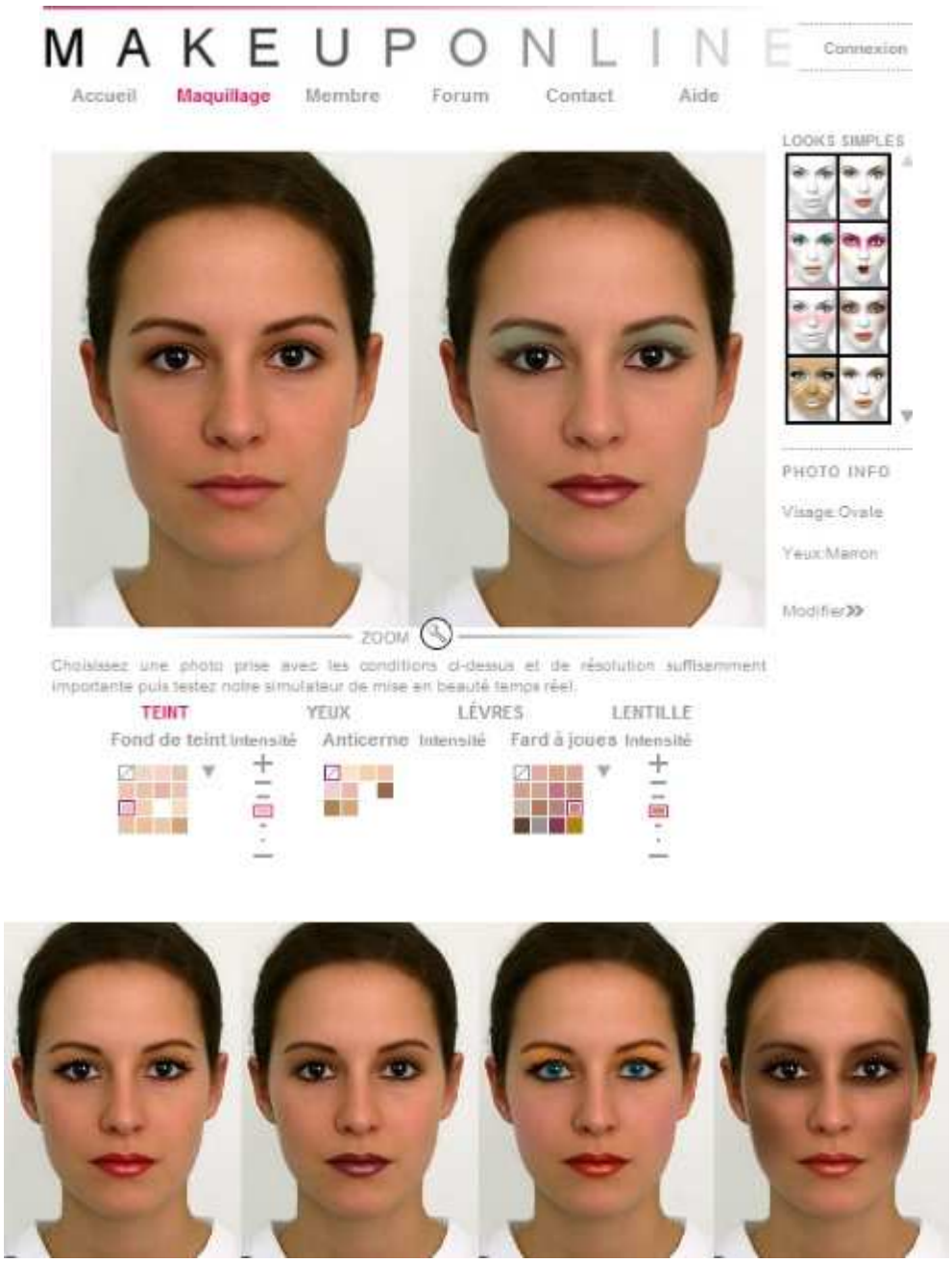


Fig. 1.03. Exemples de maquillages virtuels.

1.1.3. Améliorations et contexte du travail de thèse

La version actuelle du simulateur impose des contraintes sur l'image à modifier pour que le résultat soit satisfaisant :

- le visage doit être vu de face,
- le visage doit être suffisamment éclairé,
- la bouche doit être fermée,
- le maquillage s'applique sur une image fixe.

Le contexte du travail de thèse consiste à lever les deux dernières contraintes. Il est nécessaire de détecter le contour intérieur des lèvres, pour permettre d'appliquer du rouge à lèvre lorsqu'une personne sourit sans maquiller l'intérieur de la bouche (les dents par exemple). Et il est nécessaire de suivre les contours labiaux pour transformer esthétiquement un visage en mouvement,.
Pour répondre aux attentes du projet Makeuponline, l'objectif du travail de thèse est donc de proposer d'une part un algorithme automatique de segmentation des lèvres afin d'extraire les contours extérieur et intérieur dans des images fixes et d'autre part, de proposer un algorithme de segmentation et de suivi de ces contours afin de permettre un maquillage sur une séquence vidéo. La segmentation obtenue doit être précise pour un rendu réaliste du maquillage, robuste vis-à-vis des conditions d'acquisition et rapide.

1.2. Le projet RNTS TELMA

Le projet TELMA (**Tél**éphonie à l'usage des **mal**entendants), initié en 2006, est un projet financé par le **r**éseau **n**ational des **t**echnologies pour la **s**anté (RNTS). Ce projet a pour objectif l'étude et le développement d'un système de téléphonie multimodal accessible aux malentendants, en ajoutant des fonctionnalités audio-visuelles permettant une transcription croisée entre le code LPC (**L**angue française **P**arlée **C**omplétée) et la voix.

1.2.1. La Langue française Parlée Complétée (LPC)

La lecture sur les lèvres consiste à identifier les sons en fonction des déformations subies par la bouche. Cependant, la lecture labiale ne donne que des informations incomplètes qui, sans le son, ne peuvent être levées qu'avec le contexte de la conversation (on estime que la lecture labiale permet de percevoir seulement le tiers du message oral). Par exemple, les sons « pa », « ba » et « ma » sont produits de la même façon au niveau de la forme des lèvres; ce sont des sosies labiaux. La **L**angue française **P**arlée **C**omplétée (LPC ou code LPC) est un langage basé sur des indices visuels qui aide à lever les ambiguïtés liées à ce manque d'information.

A la différence de la **L**angue des **S**ignes **F**rançaise (LSF), le LPC n'est pas une langue à part entière. On ne parle pas en LPC, on parle en français en associant un code manuel à la parole.
Le code LPC associe des gestes des doigts et de la main à la parole. Chaque syllabe est définie par une configuration des doigts, représentant une consonne, et par une position de la main près du visage, représentant une voyelle. Il existe 8 configurations pour coder les consonnes (cf. Fig. 1.04) et 5 positions pour coder les voyelles (cf. Fig. 1.05). Le LPC fournit ainsi 40 combinaisons différentes qui permettent de lever toutes les confusions de la lecture labiale. Pris isolément, le code ne donne qu'une information partielle sur le message (à l'instar de la lecture labiale), mais la combinaison de l'image labiale et de la clé manuelle permet de visualiser la totalité du message oral.

La figure 1.06 montre un exemple du codage LPC pour la phrase suivante : « Elle a un piano à pile, et plus de pile d'ailleurs ... » (la vidéo est disponible sur le site internet de l'**A**ssociation nationale pour la promotion et le développement de la **L**angue française **P**arlée **C**omplétée http://www.alpc.asso.fr).

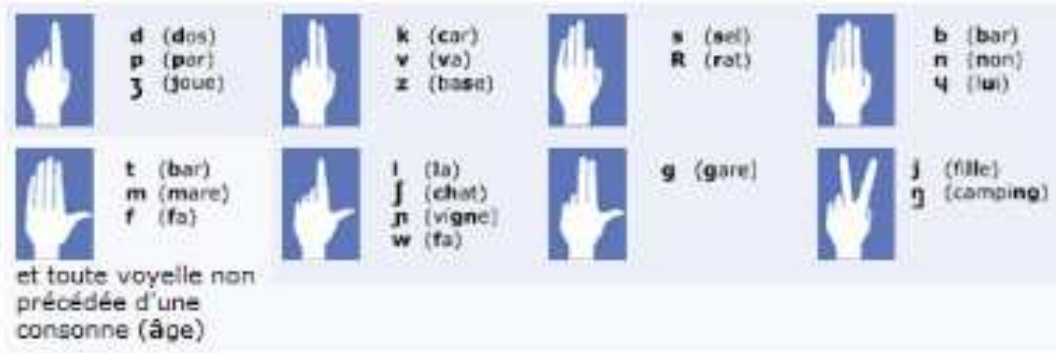


Fig. 1.04. Les 8 configurations des doigts pour coder les consonnes (source : http://www.alpc.asso.fr).

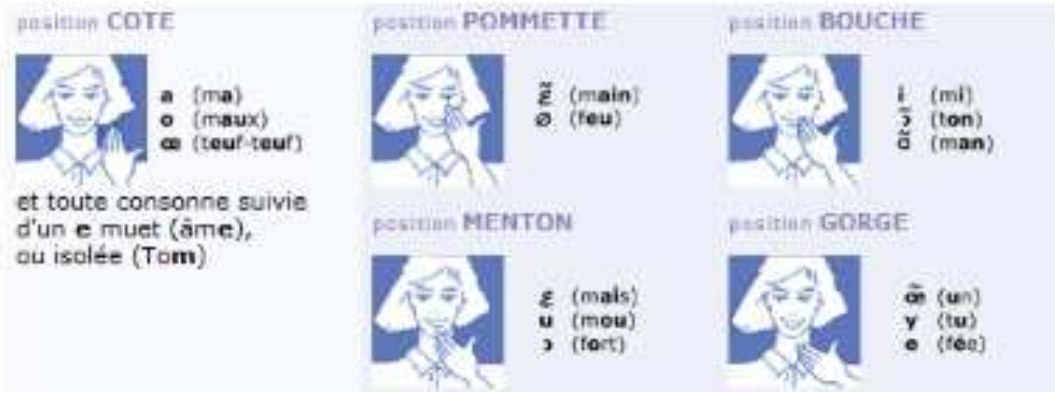


Fig. 1.05. Les 5 positions de la main pour coder les voyelles (source : http://www.alpc.asso.fr).

Fig. 1.06. Exemple de codage LPC (source : http://www.alpc.asso.fr).

1.2.2. Description du projet TELMA

Le projet TELMA a pour objectif d'utiliser le réseau téléphonique pour permettre à des personnes malentendantes de communiquer à distance. Le LPC a été choisi dans le cadre de ce projet dans la mesure où c'est un code facilement adaptable à d'autres langues (le LPC n'est pas une langue, mais un code associé à la parole) et car le LPC connaît un développement important depuis quelques années (plus de 30000 personnes francophones utilisent le LPC actuellement).

Le but du projet TELMA est de mettre en œuvre une interface multimodale permettant une conversation téléphonique dans les 3 cas de figure suivants :

- entre un locuteur bien-entendant et un locuteur bien-entendant en environnement bruité,
- entre un locuteur bien-entendant et un locuteur mal-entendant,
- entre un locuteur mal-entendant et un locuteur mal-entendant.

Les études menées au cours de ce projet concernent l'interprétation du code LPC. La figure 1.07 illustre les différentes fonctionnalités audio-visuelles du projet TELMA.

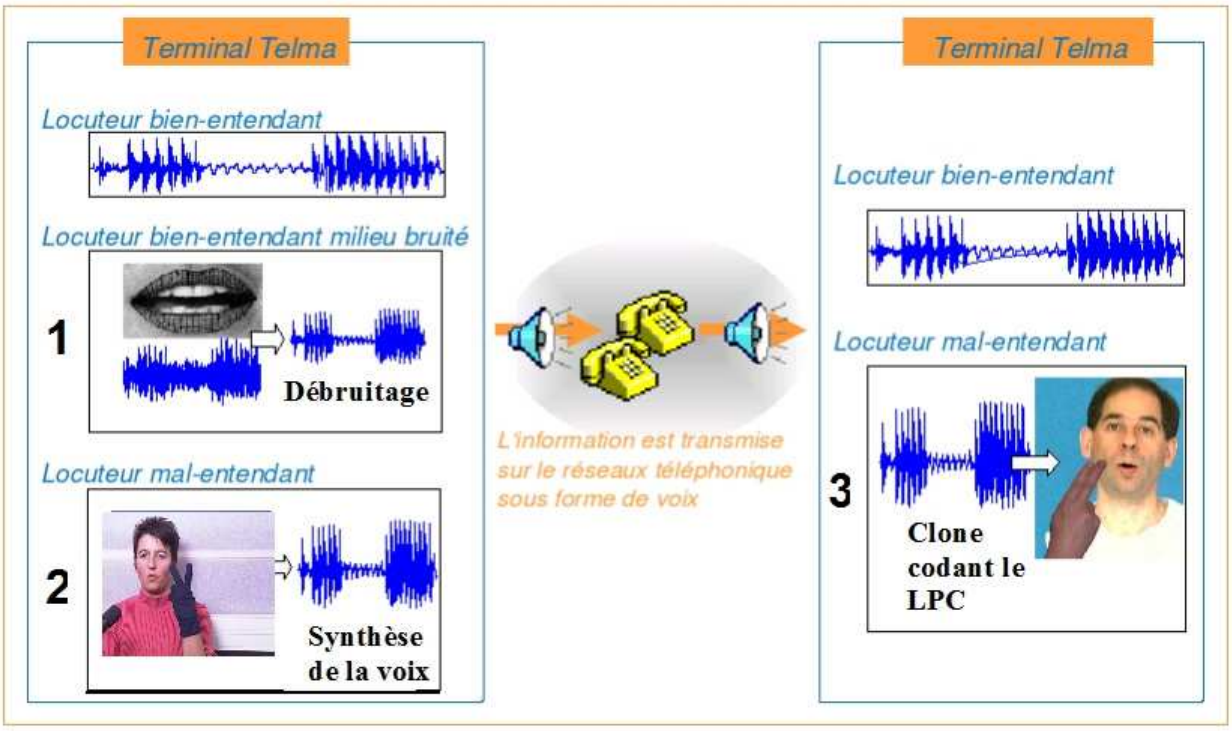


Fig. 1.07. Les fonctionnalités audio-visuelles du projet TELMA.

En plus de l'utilisation classique du réseau téléphonique, trois modules technologiques sont nécessaires :

1) **Débruitage :** Cette fonctionnalité concerne le rehaussement de la qualité de la modalité audio dans le cas d'un locuteur bien-entendant parlant dans le terminal en présence de bruit. La minimisation du bruit permet également une meilleure exploitation des restes auditifs des malentendants. Le débruitage s'appuie sur la modalité visuelle en utilisant les mouvements labiaux.

2) **Synthèse de la voix :** Un module d'interprétation des gestes manuels du LPC et l'analyse des informations extraites de la forme des lèvres permettent de produire une séquence de phonèmes

pour générer une voix synthétique destinée à être transmise par le réseau téléphonique. Cette fonctionnalité concerne le cas d'un locuteur malentendant voulant transmettre un message.

3) **Clone codant le LPC :** Dans le cas d'un locuteur malentendant recevant un message audio sur son terminal TELMA, il est nécessaire de transcrire le message oral en code LPC. Le message est décomposé en une suite de phonèmes et chacun des phonèmes est représenté par une combinaison d'une image labiale et d'une clé manuelle du LPC. Finalement, un flux vidéo synthétique est créé en animant un clone codant le LPC.

Dans le cas d'un message oral transmis sur le réseau téléphonique :

- un locuteur bien-entendant utilise soit la fonction classique du terminal, soit le module de débruitage (si l'environnement est bruité),
- un locuteur malentendant utilise le module de synthèse de la voix.

Dans le cas d'un message reçu :

- un locuteur bien-entendant écoute directement le message oral,
- un locuteur malentendant comprend le message oral à l'aide du clone codant le LPC.

1.2.3. Cadre du travail de thèse

Dans le cadre du projet TELMA, la segmentation des lèvres est une étape nécessaire pour les modules de débruitage et de synthèse de la voix.

1) Débruitage :

Concernant la partie débruitage, une étude a été menée par le **D**épartement **P**arole et **C**ognition (DPC) du GIPSA-lab dans le cadre de la thèse de B. Rivet « La bimodalité de la parole au secours de la séparation de sources » [Rivet, 2006a].
L'approche est basée sur la séparation de sources. Un signal audio inconnu $s_1(t)$ est mélangé avec d'autres signaux à partir d'une fonction de mélange $\mathcal{H}(.)$. Un critère audiovisuel utilise l'information visuelle labiale pour estimer la fonction de séparation $\mathcal{G}(.)$. L' information visuelle correspond à la forme des lèvres décrite par la largeur et la hauteur du contour intérieur de la bouche.

Lors de son travail de thèse, B. Rivet a travaillé à partir de paramètres labiaux (hauteur et largeur internes) déterminés à l'aide du système *chroma-key* [Lallouache, 1991] qui nécessite que les lèvres du locuteur soient maquillées en bleu (cf. Fig. 1.08), ce qui est très peu ergonomique.

Fig. 1.08. Conditions du système *chroma-key.*

2) Synthèse de la voix :

La figure 1.09 détaille les différentes étapes à réaliser pour produire un message oral synthétique à partir d'un locuteur malentendant codant en LPC :

- acquisition des images du locuteur malentendant,
- analyse des mouvements labiaux et reconnaissance des gestes de la main,
- fusion des deux informations visuelles,
- production de la séquence de phonèmes correspondants,
- génération du signal audio synthétique à partir de la chaîne phonétique,
- transmission de la voix synthétique sur le réseau téléphonique.

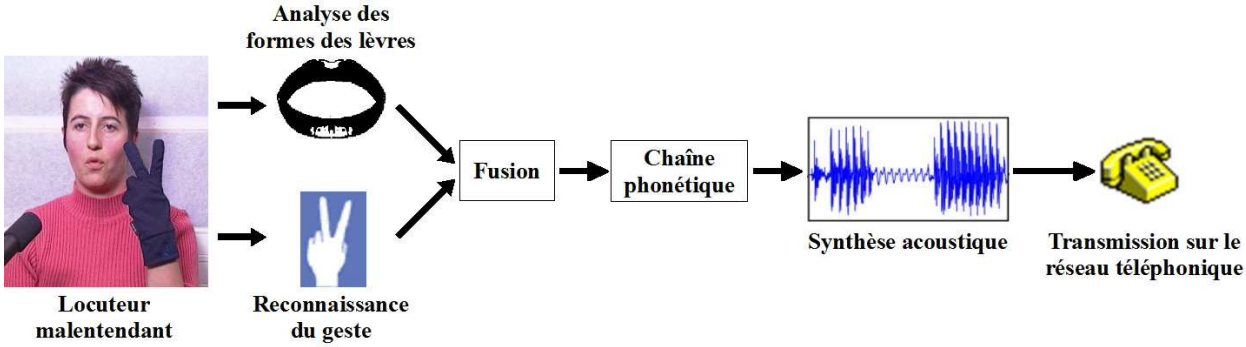


Fig. 1.09. Synthèse de la voix à partir d'un flux vidéo d'un locuteur codant en LPC.

La reconnaissance du geste de la main a été étudiée par le **D**épartement **I**mage et **S**ignal (DIS) du GIPSA-lab et France Télécom R&D dans le cadre de la thèse de Thomas Burger [Burger, 2007], « Reconnaissance automatique des gestes de la Langue française Parlée Complétée ».

La fusion des informations de l'analyse labiale et du geste de la main a été considérée dans le travail de thèse de Noureddine Aboutabit [Aboutabit, 2007a], « Reconnaissance de la Langue française Parlée Complétée (LPC) : Décodage phonétique des gestes main-lèvres », effectuée au DPC du GIPSA-lab. Dans cette étude, les deux canaux informatifs sont obtenus à l'aide d'artifices (cf. Fig. 1.10.a); un maquillage bleu pour les lèvres et le placement de pastilles de couleur sur les mains. Les contours des lèvres sont ainsi facilement segmentés par un simple seuillage sur la chrominance. L'étape de fusion nécessite de caractériser la forme de la bouche par plusieurs paramètres labiaux (cf. Fig. 1.10.b), qui sont les hauteur, largeur et aire du contour intérieur ainsi que le pincement des lèvres supérieure (*Bsup*) et inférieure (*Binf*).

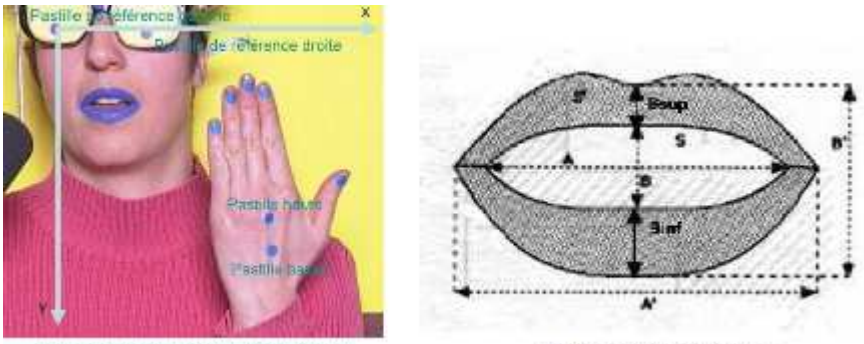


Fig. 1.10. Analyse labiale pour l'interprétation du code LPC [Aboutabit, 2007a].

L'intérêt du travail de thèse pour l'application TELMA concerne la segmentation et le suivi des contours des lèvres en vue de l'estimation des paramètres labiaux correspondants (hauteur, largeur, aire et pincement) pour s'affranchir du maquillage des lèvres en bleu.

1.3. Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (***A**utomatic **S**peech **R**ecognition ASR*) consiste à identifier des mots ou des phrases à partir d'un signal audio en comparant les sons captés avec des modèles « prototypes » préfabriqués. Actuellement, les systèmes ASR pour un petit vocabulaire (faible nombre de mots à reconnaître) et un environnement relativement contrôlé sont performants. Pour des applications réelles, les conditions de traitements (acquisition ou transmission audio) ne sont pas idéales et l'environnement bruité entraîne une baisse significative du taux de reconnaissance. Plusieurs techniques ont été proposées pour améliorer les performances en présence de bruit, telles que l'adaptation du système à l'environnement de l'application ou le débruitage du signal audio. Cependant une des solutions les plus étudiées ces dernières années consiste à utiliser une information visuelle en plus de l'acoustique pour exploiter l'aspect bimodal de la parole; on parle de reconnaissance automatique de la parole audio-visuelle (***A**udio-**V**isual **A**utomatic **S**peech **R**ecognition AV-ASR*).

1.3.1. La parole audio-visuelle

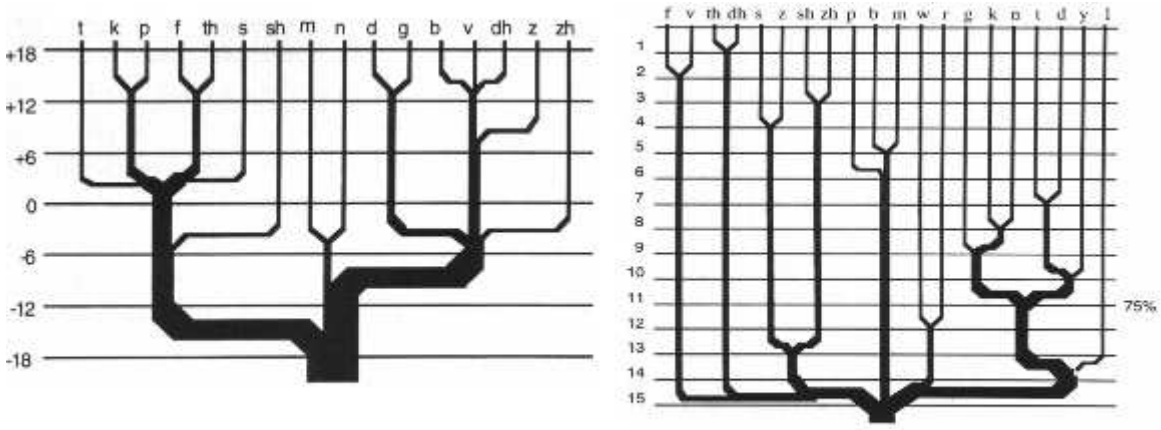
Les sciences cognitives ont depuis longtemps mis en évidence que le cerveau fusionne des informations auditives et visuelles pour mieux percevoir la parole [Neely, 1956]. Ce traitement est réalisé de manière inconsciente aussi bien en cas de mauvaises conditions acoustiques qu'en cas d'environnement peu bruité. Les personnes malentendantes utilisent la lecture sur les lèvres pour mieux comprendre une conversation, mais les personnes normo-entendantes exploitent également la bimodalité de la parole, spécialement lorsque les conditions d'intelligibilité sont mauvaises. C'est ainsi qu'au milieu de plusieurs locuteurs, nous avons tendance à nous concentrer sur les lèvres de notre interlocuteur. L'apport de l'information visuelle peut aussi être démontré dans des situations de tous les jours comme la difficulté de compréhension lors d'une discussion entre le conducteur d'une voiture et un passager assis à l'arrière, lors d'une conversation téléphonique bruitée, lors du mauvais doublage d'un film...

Cet aspect bimodal se retrouve aussi bien dans la perception que dans la production de la parole. En plus des exemples précédents, l'influence de la modalité visuelle sur la perception de la parole peut être illustrée par l'effet McGurk [McGurk, 1976]. L'effet McGurk montre que le cerveau, soumis à des sources auditives et visuelles conflictuelles, peut percevoir un son ne correspondant à aucun des deux stimuli (cf. Fig. 1.11). Par exemple, si un montage vidéo montre une personne produisant le son /ba/ mais que la source audio correspond au son /ga/, la contradiction audio-visuelle conduit généralement à percevoir le son /da/. Même si le son perçu peut toutefois être variable suivant les personnes, l'effet McGurk se produit quelque soit la langue utilisée et également avec des enfants. Dans [Easton, 1982], il est montré que l'effet inverse existe, c'est-à-dire que la perception des mouvements labiaux peut être influencée par la parole.

Stimulus audio	Stimulus visuel	Son perçu
ba	ga	da
pa	ga	ta
ma	ga	na

Fig. 1.11. Exemples de confusions dues à l'effet McGurk.

En ce qui concerne la production de la parole, en plus des cordes vocales, les sons sont émis par des articulateurs non visibles tels que la trachée, le palais, la cavité nasale ou la partie postérieure de la langue, mais aussi par des articulateurs visibles comme les lèvres, les dents, la partie antérieure de la langue ou la mâchoire. Il existe donc une complémentarité entre la parole produite et la parole visuelle. Cette complémentarité permet de distinguer visuellement des phonèmes proches auditivement ou inversement. Dans [Summerfield, 1987], des arbres de confusion auditive et visuelle des consonnes sont construits expérimentalement en présentant des stimuli à des adultes normo-entendants et en classant les confusions en fonction du bruit (cf. Fig. 1.12). Nous pouvons remarquer que les consonnes /k/ et /p/ sont ambigües avec l'audio mais facilement reconnaissable visuellement. De même, /f/ et /v/ sont proches visuellement et éloignées auditivement.



a. Confusion auditive. L'échelle verticale correspond au RSB (en dB)

b. Confusion visuelle. L'échelle verticale correspond au niveau de regroupement

Fig. 1.12. Arbres de confusion des consonnes [Summerfield, 1987].

De la même manière que le cerveau intègre des informations audio-visuelles et afin de tirer profit de la complémentarité entre les deux modalités, les systèmes AV-ASR utilisent le traitement d'images pour augmenter le taux de reconnaissance de la parole par rapport à un système ASR n'utilisant que le canal auditif.

1.3.2. Systèmes de reconnaissance de la parole audio-visuelle

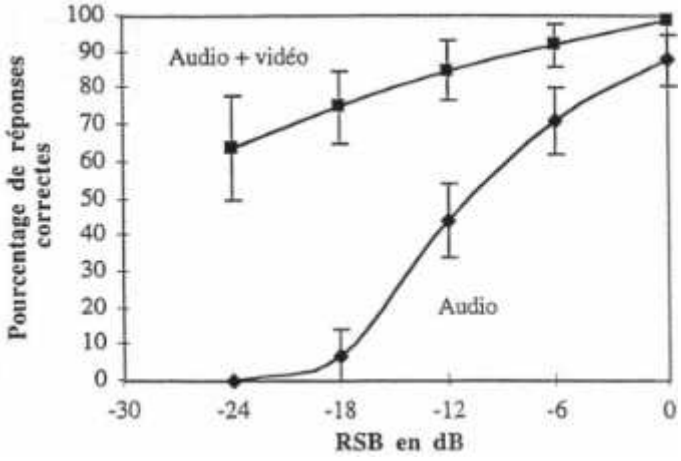


Fig. 1.13. Apport de l'information visuelle en reconnaissance de la parole [Benoît, 1994].

Déjà en 1954, Sumby *et al.* [Sumby, 1954] ont montré les bénéfices que pouvait apporter la modalité visuelle pour augmenter le taux de reconnaissance de la parole en milieu bruité. Dans [Benoît, 1994], le taux de reconnaissance en n'utilisant que le canal audio est comparé avec l'apport de la vision. La figure 1.13 présente les résultats de l'étude et le pourcentage de réponses correctes en fonction du bruit. Ces résultats montrent que les performances du système sont améliorées même lorsque le bruit est faible. En conséquence, les études sur la reconnaissance de parole ont très vite proposé d'intégrer un traitement visuel. Le premier système AV-ASR a été mis en œuvre par Petajan en 1984 [Petajan, 1984]. Depuis, ils ont tous le même schéma de principe (cf. Fig. 1.14) et le processus est le suivant :

- acquisition du signal audio-visuel réalisée par un ou plusieurs microphones et une caméra,
- extraction de paramètres audio et visuels,
- fusion audio-visuelle des données,
- reconnaissance de la parole.

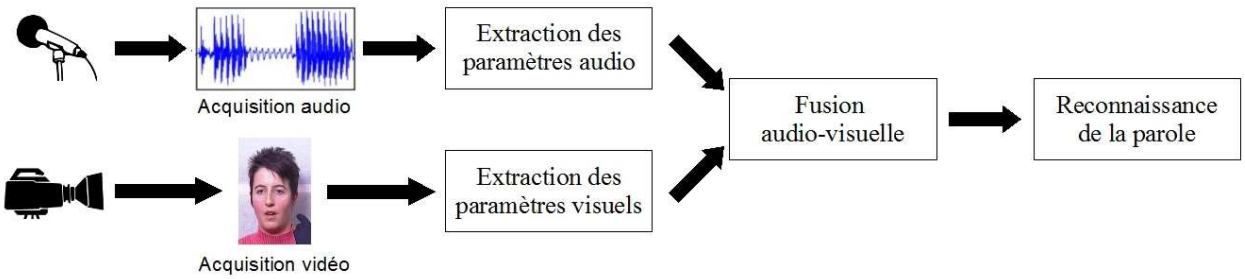


Fig. 1.14. Schéma de principe d'un système AV-ASR.

Classiquement, on distingue deux types de fusion audio-visuelle : la fusion des paramètres ou la fusion des scores. La fusion des paramètres est réalisée dès que les paramètres des signaux audio et vidéo sont extraits. La fusion des scores est utilisée lorsqu'on sépare les systèmes audio et vidéo. Chacun des deux systèmes propose une décision et ce sont ces décisions qui sont combinées; des poids différents peuvent être attribués pour privilégier une des deux modalités.
En ce qui nous concerne, nous ne détaillerons pas les parties correspondantes à l'extraction des paramètres audio, à la fusion des données et à l'étape de reconnaissance. Le lecteur peut se référer à l'étude [Potamianos, 2004] pour avoir un aperçu plus complet de la reconnaissance audio-visuelle. Nous nous intéressons à l'information visuelle utilisée par les systèmes AV-ASR.

1.3.2.a. Information visuelle utile

Pour mieux comprendre un message oral, il est possible d'observer le comportement ou les gestes d'une personne, mais la source principale d'information est sans conteste le visage. Mais quelles sont les caractéristiques faciales les plus informatives?

Nous avons montré que la production de parole était bimodale et qu'elle était réalisée notamment par des articulateurs visibles tels que la mâchoire, les lèvres, les dents ou la partie antérieure de la langue. La figure 1.15 montre la différence du taux de reconnaissance en fonction du bruit suivant que seul l'audio, l'audio et les lèvres ou l'audio et le visage sont utilisés. Cette étude menée par Benoît *et al.* [Benoît, 1996] montre que les lèvres véhiculent plus des deux tiers de l'information visuelle contenue dans le visage. Ainsi, même si d'autres parties du visage sont mis en jeu lors de la production, les mouvements et la forme des lèvres portent la majeure partie de l'information visuelle. En conséquence, les paramètres visuels employés par les systèmes AV-ASR sont généralement extraits à partir de la bouche.

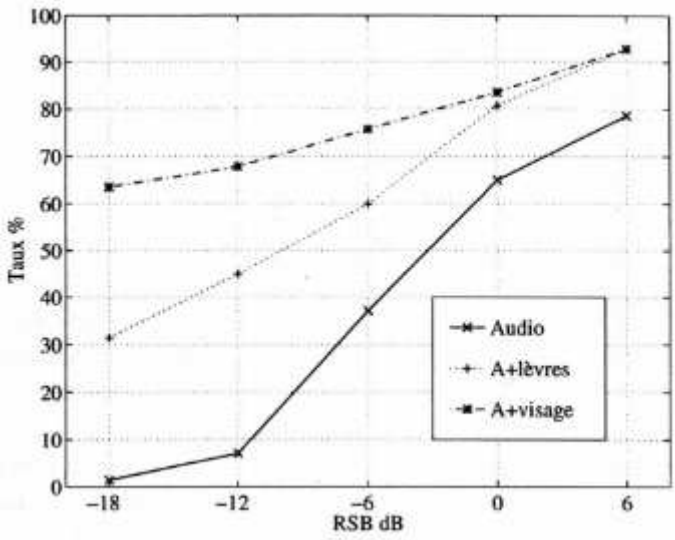


Fig. 1.15. Intelligibilité de la parole en fonction des éléments visuels disponibles [Benoît, 1996].

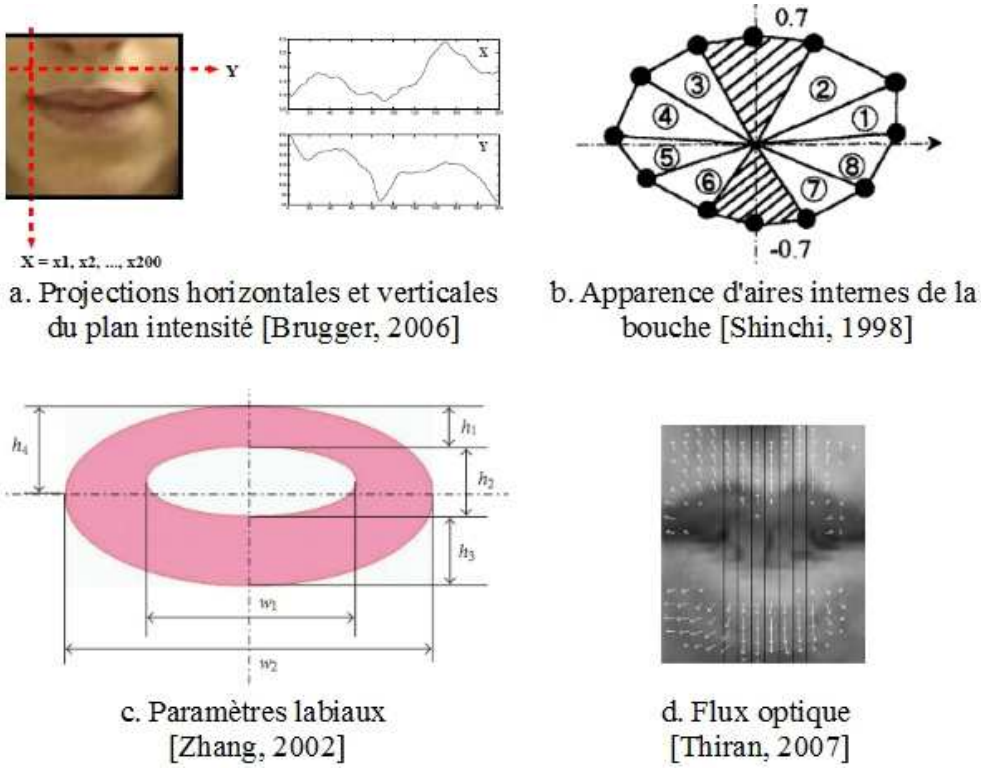
1.3.2.b. Détection des paramètres visuels à partir de la bouche

Au cours des 25 dernières années, plusieurs caractéristiques labiales ont été testées en vue d'être intégrées aux systèmes de communication audio-visuels. De manière générale, les paramètres sont extraits à partir de l'apparence de la bouche, de la forme de la bouche (contours des lèvres) ou en combinant ces deux types d'information.

Les approches basées sur l'apparence suggèrent que tous les pixels de la région d'intérêt sont porteurs d'information. Une démarche classique consiste à utiliser la Transformée en Cosinus Discret (***D**iscrete **C**osine **T**ransform DCT*) comme dans [Potamianos, 1998]. Ceci s'explique par le fait que la DCT possède une excellente propriété de regroupement de l'énergie du signal d'entrée sur un faible nombre de coefficients DCT et qu'elle peut être implémentée rapidement d'une manière similaire à la Transformée de Fourier rapide. Dans [Heckmann, 2002], Heckmann *et al.* sélectionnent les coefficients DCT d'énergie et de variance importantes calculés à partir de la région de la bouche. Brugger *et al.* adoptent une autre approche basée sur les projections [Brugger, 2006]. Une fois la zone de la bouche localisée, les dimensions de la région d'intérêt sont normalisées (taille de 200 x 200 pixels) et les projections sur l'axe horizontal (resp. vertical) sont calculées comme la somme des niveaux de gris sur chaque colonne (resp. chaque ligne) (Fig. 1.16.a). Dans [Shinchi, 1998], la reconnaissance est effectuée en analysant 8 aires en forme de triangle construites en reliant le point situé au centre du contour extérieur avec 12 points clefs situés sur le contour (cf. Fig. 1.16.b). Yuhas *et al.* [Yuhas, 1989] proposent de mettre directement les valeurs de luminance des pixels de la bouche en entrée d'un réseau de neurones.

Les méthodes utilisant les formes des lèvres considèrent que l'information visuelle utile est surtout présente dans les contours labiaux. Dans [Chan, 1998] et [Barnard, 2002], les paramètres labiaux sont la hauteur et la largeur du contour extérieur des lèvres. Finn *et al.* [Finn, 1988] calculent 14 distances caractéristiques à partir du contour extérieur. Il est toutefois plus efficace d'utiliser également des paramètres calculés à partir du contour intérieur. Dans [Petajan, 1984], les caractéristiques visuelles employées pour la reconnaissance sont les hauteur, largeur, surface et périmètre de la région interne de la bouche. Dans [Lallouache, 1991] et [Zhang, 2002], les paramètres labiaux sont définis à la fois à partir du contour extérieur et à partir du contour intérieur (cf. Fig. 1.16.c). L'évolution temporelle des paramètres labiaux a également été considérée dans [Nishida, 1986] et [Goldshen, 1993] en calculant la dérivée temporelle des paramètres. Dans le même genre d'idée, les mouvements des lèvres ont été exploités dans [Mase, 1991] et [Thiran, 2007] en utilisant le flux optique. Thiran *et al.* utilisent l'algorithme de flux optique de Lucas-Kanade [Lucas, 1981] et les paramètres visuels correspondent aux mouvements relatifs

moyens horizontaux et verticaux dans la région de la bouche (cf. Fig. 1.16.d); un paramètre supplémentaire est l'intensité du point central de la bouche.



a. Projections horizontales et verticales du plan intensité [Brugger, 2006]

b. Apparence d'aires internes de la bouche [Shinchi, 1998]

c. Paramètres labiaux [Zhang, 2002]

d. Flux optique [Thiran, 2007]

Fig. 1.16. Exemple de définition de paramètres visuels utilisés en reconnaissance de parole.

Certaines études fusionnent des informations globales et locales. Dans [Potamianos, 1997], la transformée en ondelettes de la région de la bouche est combinée avec la hauteur et la largeur du contour extérieur. Seguier *et al.* [Seguier, 2003] utilisent la hauteur et la largeur de la bouche, leurs dérivées temporelles, ainsi que le pourcentage de pixels clairs et sombres à l'intérieur de la bouche.
En plus des lèvres, d'autres parties visibles du visage peuvent être utilisées en tant que paramètres visuels comme l'indication de présence des dents et de la langue ([Chan, 2001]; [Zhang, 2002]). Ces études montrent que les taux de reconnaissance de la parole sont plus performants en ajoutant ces deux informations.

1.4. Autres applications de l'analyse labiale

1.4.1. Exploitation de la parole audio-visuelle

La reconnaissance automatique de la parole a été le premier thème de recherche à exploiter l'aspect bimodal de la parole, mais nous pouvons brièvement présenter d'autres applications utilisant cette particularité.

- *Rehaussement de la qualité audio en environnement bruité :*

Dans [Girin, 1997], Girin *et al.* présente un système de débruitage de la parole en intégrant des informations auditives et visuelles.

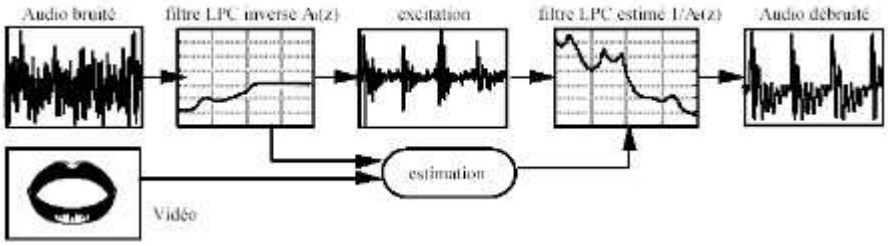


Fig. 1.17. Schéma de principe du système de débruitage introduit dans [Girin, 1997].

Comme illustré sur la figure 1.17, le système utilise le modèle de prédiction linéaire et les filtres LPC (***L**inear **P**redictive **C**oding*). Les paramètres labiaux définissant l'information visuelle sont les hauteur, largeur et aire interolabiales.
L'idée d'estimer des filtres de rehaussement par une prédiction linéaire des paramètres audio à partir des paramètres vidéo a été reprise dans des systèmes de reconnaissance de la parole dans ([Deligne, 2002]; [Goecke, 2002]).

- *Séparation de sources audio-visuelle :*

La séparation de sources audio-visuelle peut être vue comme une extension du problème de rehaussement lorsque plusieurs sources audio bruitées sont enregistrées par plusieurs capteurs. Le principe consiste à estimer à partir de l'image du locuteur et du mouvement des lèvres un modèle statistique audio-visuel du signal audio prononcé, puis à filtrer le signal audio bruité grâce au modèle estimé ([Sodoyer, 2004]; [Wang, 2005]).

- *Détection de l'activité vocale (**V**oice **A**ctivity **D**etection VAD) :*

Les système VAD visuels utilisent le signal vidéo pour détecter l'activité de parole. L'idée est d'exploiter le fait que les lèvres bougent durant la production de parole, alors que le mouvement des lèvres est inexistant ou faible pendant les silences. Dans [Rivet, 2006b], les dérivées temporelles de deux paramètres labiaux (hauteur et largeur du contour intérieur) sont calculées et la classification silence/parole est basée sur un seuillage de ces valeurs. Dans [Rivet, 2007], l'information visuelle est plus globale et deux méthodes l'une basée sur les **M**odèles **A**ctifs d'**A**pparence (AAM) et l'autre sur un filtrage rétinien sont comparées.

- *Compression audio-visuelle :*

Ce type de compression exploite la redondance et la complémentarité existant entre le signal audio et le signal vidéo pour coder conjointement les deux signaux. Comparée aux systèmes classiques où les signaux sont codés séparément, la compression audio-visuelle permet de diminuer le débit de transmission et la complexité des codeurs. Ceci est particulièrement utile pour des applications de visiophonie par exemple. Girin introduit un système de compression audio-visuelle dans [Girin, 2004] où les paramètres visuels sont extraits à partir des contours extérieur et intérieur des lèvres.

1.4.2. Animation de tête parlante

Après la reconnaissance automatique de la parole, le deuxième thème de recherche le plus étudié concerne l'animation virtuelle du visage. Pour montrer l'intérêt d'utiliser des clones dans les systèmes de communication audio-visuels, Beskow *et al.* [Beskow, 1997] étudient l'intelligibilité de la parole en fonction de signaux audio et vidéo naturels et synthétiques (cf. Fig. 1.18).

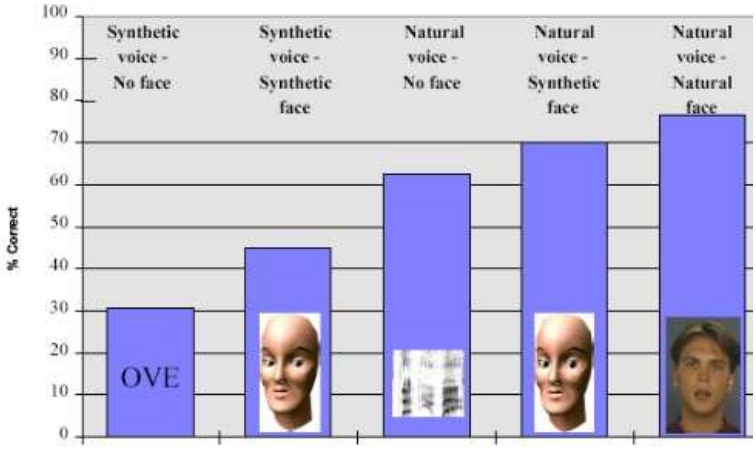


Fig. 1.18. Test d'intelligibilité [Beskow, 1997].

Le test d'intelligibilité consiste, pour des sujets adultes, à reconnaître des enchaînements VCV (Voyelle-Consonne-Voyelle) en environnement bruité (RSB de 3 dB). L'étude propose différentes sources pour le signal audio-visuel : 1) une voix synthétique et pas de signal vidéo, 2) une voix et un visage synthétiques, 3) une voix naturelle et pas de signal vidéo, 4) une voix naturelle et un visage synthétique et 5) une voix et un visage naturels. Les résultats présentés sur la figure 1.18 montrent que le taux de reconnaissance en utilisant le visage synthétique est proche des conditions naturelles.

L'avantage principal pour les systèmes de communication est que la transmission de paramètres contrôlant un visage virtuel coûte moins chère en terme de débit que de transmettre le signal vidéo. Les avatars peuvent également servir dans des applications d'interaction homme-machine pour rendre l'interface entre l'utilisateur et l'ordinateur plus naturelle. Les déformations du visage, dont celles des lèvres, ont été modélisées par les Unités d'action et les *FFP* et *FAPU* du standard MPEG-4. Les visages synthétiques sont généralement contrôlés par ce genre de descripteurs.

- Unités d'action (*Action Units*) : En 1978, Ekman *et al.* [Ekman, 1978] proposent un système de codage manuel des expressions faciales. Les mouvements élémentaires des muscles sont décomposés en 46 *Action Units* permettant de décrire tous les mouvements visibles du visage. Chaque mouvement peut être représenté par la combinaison d'une ou plusieurs *Action Units*.

Fig. 1.19. Exemple d'*Action Units* pour le bas du visage.

- *FDP* et *FAPU* de MPEG-4 : Le standard MPEG-4 [MPEG, 2001] possède un modèle 3D articulé du visage. L'animation du modèle utilise des points caractéristiques du visage et des *Animation Units*.

  - *Facial Definition Parameters* (FDP) : Les FDP sont un ensemble de points caractéristiques du visage (cf. Fig. 1.20 pour les FDP de la bouche). Ils permettent de personnaliser l'animation en représentant la topographie ou l'ossature d'un visage particulier.

  - *Facial Animation Parameters Units* (FAPU) : Les FAPU permettent de définir les mouvements élémentaires du visage (équivalents des *Action Units* introduits par Ekman *et al.* [Ekman, 1978]).

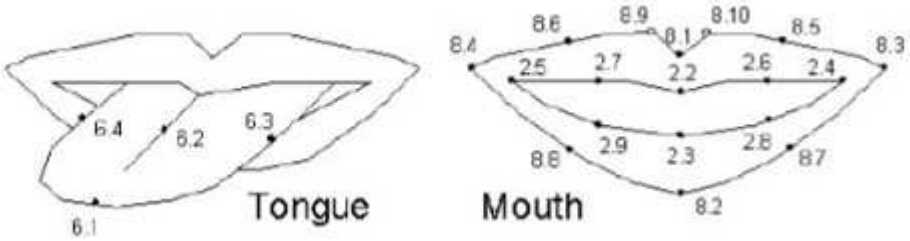


Fig. 1.20. Points caractéristiques *FDP* utilisés pour décrire la bouche dans MPEG-4.

Les *Action Units* et les paramètres FDP et FAPU sont fréquemment utilisés pour animer des visages virtuels. Par exemple, CANDIDE [Ahlberg, 2001] est un masque paramétrique du visage (cf. Fig.1.21) disponible en 3 versions. Candide-1 et Candide-2 utilisent respectivement 11 et 6 *Action Units*, alors que Candide-3 est animé à partir des FAPU.

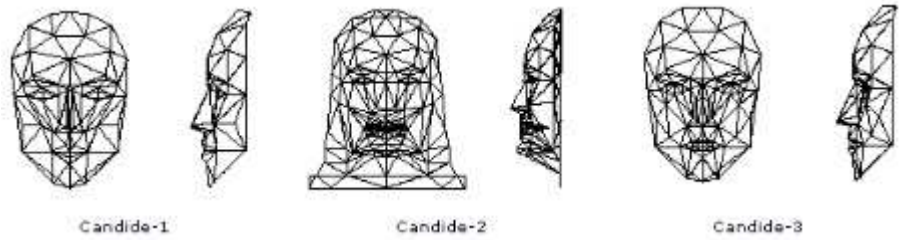
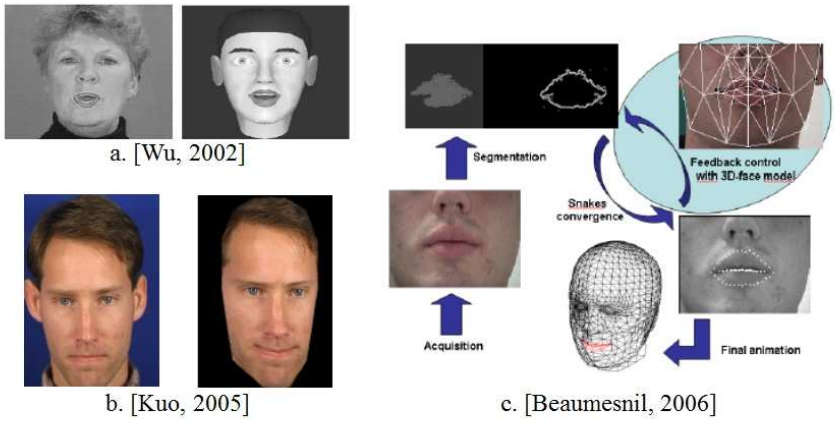


Fig. 1.21. Les 3 versions du masque virtuel du visage CANDIDE [Ahlberg, 2001].

Nous pouvons mentionner différentes études de la littérature qui visent à animer des têtes parlantes. Dans [Morishima, 1989], les auteurs animent un modèle 3D du visage en utilisant une analyse cepstrale du signal audio. Ensuite, les avancements en traitement d'images ont permis d'utiliser le signal vidéo.

Dans ([Terzopoulos, 1993]; [Botino, 2002]; [Wu, 2002]; [Yin, 2002]; [Kuo, 2005]; [Beaumesnil, 2006]), les paramètres d'animation sont obtenus en détectant les contours des lèvres (approche locale). Dans [Wu, 2002], les FAPU de MPEG-4 sont déterminés à partir des contours labiaux extérieur et intérieur (cf. Fig. 1.22.a). Kuo *et al.* [Kuo, 2005] détectent le contour extérieur des lèvres pour des images de bouches fermées afin d'animer la bouche virtuelle avec une seule *Action Unit* représentant le sourire (cf. Fig. 1.22.b). Dans [Beaumesnil, 2006], les auteurs corrigent la détection des contours labiaux à l'aide du modèle Candide-3 pour obtenir des paramètres d'animation plus précis afin d'animer un modèle 3D plus complexe (cf. Fig. 1.22.c).

Des informations visuelles globales peuvent être également utilisées pour animer les clones. Par exemple, Essa *et al.* [Essa, 1994] se servent du flux optique calculé sur l'ensemble du visage. Les Modèles Actifs d'Apparence sont également beaucoup utilisés pour obtenir les paramètres MPEG-4 ([Lehn-Schiøler, 2004]; [Abboud, 2005]).



a. [Wu, 2002]

b. [Kuo, 2005]

c. [Beaumesnil, 2006]

Fig. 1.22. Exemples d'animation virtuelle du visage.

1.4.3. Reconnaissance d'expressions du visage

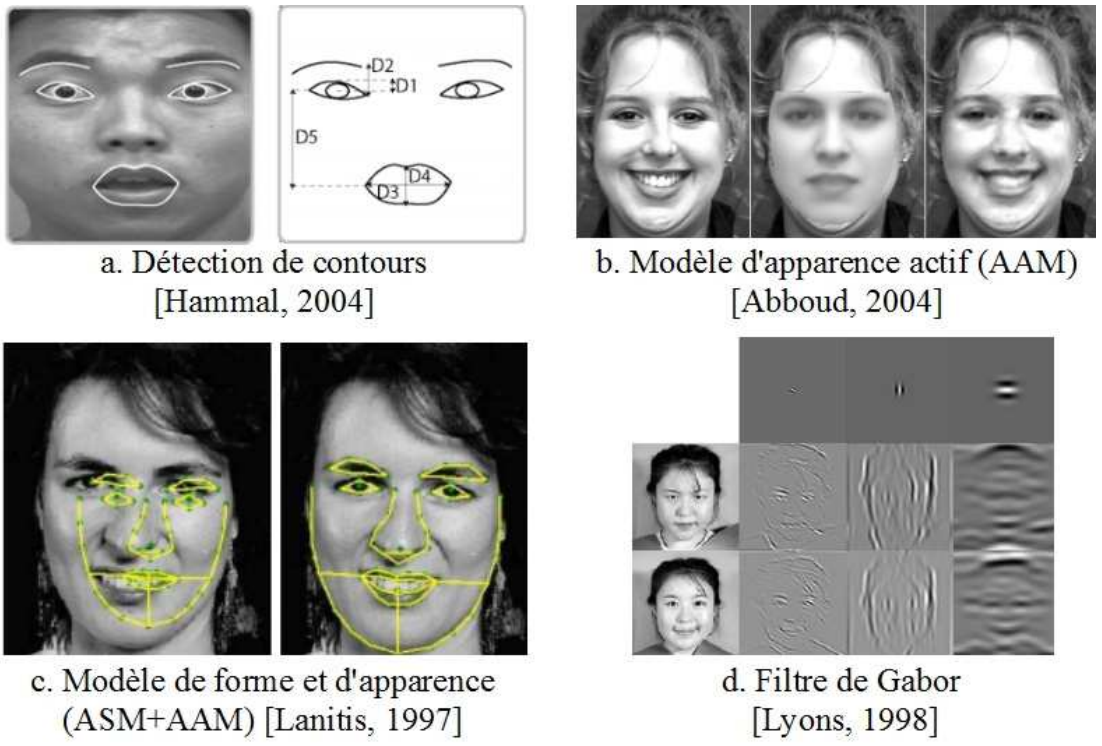
Déjà étudiées au XIX^{e} siècle par Darwin et Duchenne, les expressions du visage sont un référent puissant du comportement et de la communication humaine non-verbale. Les expressions faciales permettent d'obtenir des informations sur les émotions ressenties ou les intentions d'une personne. Elles se caractérisent par la modification de l'aspect du visage due à l'activation d'un ou plusieurs muscles faciaux. Par exemple, en observant le visage, il est aisé de reconnaître la tristesse (coins des yeux et de la bouche dirigés vers le bas), la joie (apparition du sourire et yeux qui se referment) ou la surprise (sourcils dirigés vers le haut et bouche qui s'ouvre). En outre, les expressions du visage jouent un rôle important dans la langue des signes, puisqu'elles s'apparentent aux intonations de la voix chez les normo-entendants.

Dans le passé, les expressions étaient analysées principalement dans le cadre de la psychologie ou de la biologie. Aujourd'hui avec les progrès réalisés en traitement d'images et en analyse du visage par ordinateur, les applications sont multiples et touchent essentiellement les domaines de l'interaction homme-machine (pour que l'ordinateur réagisse en fonction du comportement de l'utilisateur), de la reconnaissance du visage (pour développer un système invariant aux expressions), de l'animation d'avatars (pour une animation plus réaliste, par exemple MPEG-4 intègre un niveau de description des expressions pour son modèle 3D du visage, tels que la colère, le dégoût, la joie, la tristesse, la peur, la surprise et l'état neutre)...

Les méthodes de reconnaissance d'expressions faciales nécessitent une première étape qui consiste à extraire les informations pertinentes du visage. Ces informations sont obtenues de deux manières, soit par une approche globale (modélisation du visage entier), soit par une approche locale (détection de traits caractéristiques). L'interprétation des informations faciales extraites permet d'identifier les différentes expressions. On distingue deux méthodes de classification fréquemment employées : soit on utilise directement les informations visuelles détectées (information globale ou contours) et la reconnaissance peut être effectuée par un réseau de neurones ou un modèle de Markov caché (***H**idden **M**arkov **M**odel HMM*) par exemple, soit on détecte les *Action Units* présentes sur le visage. Nous avons vu que les *Action Units* introduites par Ekman *et al.* [Ekman, 1978] permettent de décrire les déformations faciales, ainsi les expressions du visage peuvent être définies par une combinaison d'une ou plusieurs *Action Units* parmi les 46 existantes. Des études telles que ([Yamada, 1993]; [Morishima,

1995]) ont montré que les parties du visage les plus informatives pour la classification des expressions sont les yeux et la bouche. De manière générale, les systèmes de reconnaissance automatique des expressions faciales cherchent à reconnaître une des 6 expressions universelles introduites par Ekman [Ekman, 1982] : la colère, le dégoût, la joie, la tristesse, la peur et la surprise; une septième « expression » peut être l'état neutre.

Dans [Seyedarabi, 2006], les contours extérieurs de la bouche sont suivis et des points caractéristiques du contour permettent d'identifier des *Actions Units* du bas du visage à l'aide d'un réseau de neurones. Le même type d'étude est réalisé dans [Tian, 2001]; les traits permanents des yeux, de la bouche, des sourcils, des joues et des rides près du nez servent à reconnaître 7 *Action Units* du haut du visage et 11 autres pour le bas. Dans [Hammal, 2007], les contours de la bouche, des yeux et des sourcils sont extraits afin de calculer 5 distances caractéristiques (cf. Fig. 1.23.a). La théorie de l'évidence permet de reconnaître les expressions universelles à partir de l'évolution temporelle de ces distances. Olivier *et al.* [Olivier, 2000] ne se servent que du contour extérieur précis de la bouche pour la classification d'expressions. On peut également mentionner des exemples d'études utilisant une approche plus globale pour extraire les informations du visage. Dans ([Yacoob, 1994]; [Cohn, 1998]; [Tian, 2000b]), le flux optique est utilisé pour estimer le mouvement des différentes parties du visage. Nkambou *et al.* [Nkambou, 2004] utilisent la décomposition du visage en « eigenfaces » et un réseau de neurones pour identifier les expressions. Les « eigenfaces » permettent de décomposer l'image dans un sous-espace réduit; elles sont également étudiées dans [Turk, 1991]. Dans [Abboud, 2004], un **M**odèle **A**ctif d'**A**pparence (AAM) est construit à partir de 375 images de visage annotées manuellement pour extraire des vecteurs de texture. Ce modèle s'adapte de manière itérative sur une image inconnue (cf. Fig. 1.23.b) et les vecteurs de paramètres d'apparence associés à la déformation du modèle permettent de reconnaître l'expression faciale à l'aide d'une régression linéaire. Les AAM ont largement été utilisés pour la reconnaissance des expressions du visage ([Hong, 2006]; [Ratliff, 2008]). Un AAM est associé à un Modèle Actif de Forme (ASM) dans [Lanitis, 1997] pour obtenir quelques traits caractéristiques du visage (cf. Fig. 1.23.c). Dans ([Lyons, 1998]; [Yang, 2005]; [Liu, 2006]), les caractéristiques des mouvements faciaux sont obtenues à l'aide des filtres de Gabor (cf. Fig. 1.23.d).



a. Détection de contours [Hammal, 2004]

b. Modèle d'apparence actif (AAM) [Abboud, 2004]

c. Modèle de forme et d'apparence (ASM+AAM) [Lanitis, 1997]

d. Filtre de Gabor [Lyons, 1998]

Fig. 1.23. Exemple d'extraction d'indices faciaux pour la reconnaissance d'expressions.

1.4.4. Identification de personnes

Ces dernières années, les techniques d'identification biométriques se sont multipliées pour répondre à un besoin croissant dans le domaine de la sécurité. La biométrie permet l'identification des personnes en fonction de caractéristiques biologiques (ADN, sang, salive...), comportementales (voix, dynamique de la signature, démarche...) et morphologiques (empreintes digitales, traits du visage, iris...). Le domaine de la vision par ordinateur s'est concentré sur le dernier aspect en proposant, en particulier, des méthodes de reconnaissance et d'analyse du visage.

De manière générale, on distingue deux types d'identification suivant l'application visée :

- la reconnaissance d'individus : elle permet de retrouver l'identité d'une personne
- l'authentification de l'utilisateur : elle est la preuve et elle sert à confirmer une identité.

La reconnaissance concerne principalement la sécurité et la télésurveillance. Elle peut être utilisée par la police ou les services d'immigration dans des lieux publics (aéroports, gares, stade...) pour identifier des personnes recherchées appartenant à une base de données. L'authentification est employée dans les domaines nécessitant un contrôle d'accès (lieux sécurisés, systèmes bancaires, internet...).

La biométrie faciale peut sembler être une méthode peu sûre dans la mesure où elle est sujette à des variations élevées (expression, déguisement, barbe, lunette, vieillissement, conditions d'éclairage...). Mais elle reste un thème extrêmement étudié car elle permet une acquisition aisée des informations (caméra) pour la reconnaissance et il est possible de combiner des éléments morphologiques avec des éléments comportementaux pour l'authentification (forme de la bouche et voix par exemple). Ces avantages sont intéressants comparés à des techniques basées sur les empreintes digitales et l'analyse de l'iris qui, même si elles sont très sûres, demandent des périphériques spécifiques, coûteux et contraignants pour scanner l'extrémité du doigt et le détail de l'œil.

Parmi les composantes faciales étudiées dans les travaux portant sur la reconnaissance et l'authentification des personnes, la bouche a été particulièrement analysée. De la même manière que pour les empreintes digitales, les empreintes des lèvres sont connues en médecine pour être spécifiques à chaque personne [Suzuki, 1970]. En outre, la forme de la bouche et les déformations labiales associées à un message oral varient beaucoup d'un individu à l'autre. Les techniques d'analyse labiale proposent de comparer les contours labiaux seuls ou de les associer à la voix pour une reconnaissance audio-visuelle.

Dans [Wark, 1998], les contours extérieurs des lèvres sont suivis sur des séquences d'images. Les caractéristiques labiales extraites sont les profils chromatiques *R, G, B* le long des normales des points du contour. Une analyse en composantes principales permet de réduire la dimension des données et la classification est réalisée à l'aide d'un modèle de mélange de gaussiennes. Dans [Luettin, 1996], les contours extérieur et intérieur des lèvres sont obtenus à partir d'un ASM et l'identification est réalisée avec un modèle de Markov caché.

Dans [Brand, 2001], un système de reconnaissance labiale est combiné avec un système de reconnaissance de la parole. Les auteurs montrent que les contours des lèvres doivent être très précis; ainsi pour obtenir les paramètres labiaux, ils maquillent les lèvres en bleu et ils utilisent le système *Chroma-key* [Lallouache, 1991]. L'étude montre que les taux de reconnaissance du système hybride sont plus élevés que ceux obtenus en prenant séparément les caractéristiques visuelles et auditives. Jourlin *et al.* [Jourlin, 1997] développent également le même genre de système hybride. Les formes des lèvres sont obtenues automatiquement à l'aide d'un ASM et l'étude montre qu'un poids supérieur doit être affecté aux données issues de l'audio pour obtenir un meilleur taux de reconnaissance.

1.4.5. Application médicale

Les progrès importants qu'a connu l'imagerie médicale ces vingt dernières années ont amené à développer des outils performants de traitement d'images, notamment de la segmentation, pour des applications médicales. Si ce domaine concerne principalement l'imagerie radiologique (segmentation de tumeurs cancéreuses, des parties du cerveau, des poumons...), quelques études proposent également d'extraire des caractéristiques visuelles de la bouche pour l'aide médicale assistée par ordinateur.

Les informations extraites concernent les lèvres, la langue ou les dents. Dans [Yokogawa, 2007], Yokogawa *et al.* détectent précisément le contour extérieur de la bouche afin d'organiser la procédure de chirurgie plastique des lèvres. La méthode est également utilisée pour des applications dentaires. Zou *et al.* [Zou, 2007] réalisent l'extraction des contours de la langue pour obtenir un diagnostic médical automatique.

1.5. Conclusion

Dans ce premier chapitre, nous avons présenté différentes applications nécessitant d'effectuer une analyse de la zone des lèvres. Les déformations caractéristiques des contours de la bouche servent à développer des systèmes de reconnaissance automatique de la parole performants, à animer des avatars de manière réaliste ou à reconnaître des émotions. La spécificité de l'empreinte des lèvres est exploitée pour faire de la reconnaissance de personnes et pour des applications médicales. La grande diversité de ces thèmes de recherche mettent en évidence l'utilité de l'analyse labiale.
Nous avons également montré que les informations visuelles utiles sont variables selon l'application visée. Elles peuvent être obtenues par une approche globale (qui tient compte de toute la zone de la bouche) ou par une approche locale (détection des contours des lèvres).

Dans le cadre de cette thèse, les algorithmes proposés pour segmenter les lèvres doivent prendre en considération les contraintes de nos deux applications cibles : le produit Makeuponline de Vesalis et le projet TELMA. L'extraction des contours externe et interne doit être précise et robuste aux conditions d'illumination pour appliquer un maquillage réaliste avec le logiciel Makeuponline. Le suivi temporel des contours doit fournir des paramètres labiaux (hauteur, largeur, aire de la zone interne et pincements) pour le projet TELMA.

CHAPITRE 2

Etat de l'art : espaces couleurs et segmentation des lèvres

L'extraction des contours des lèvres reste une tâche ardue et différentes approches ont été proposées dans la littérature lors des vingt dernières années sans résoudre complètement le problème. La difficulté de la segmentation s'explique par plusieurs raisons :

- La bouche est une composante faciale hautement déformable (cf. Fig. 2.01). Suivant l'ouverture de la bouche (fermée, ouverte), les contours des lèvres varient beaucoup. La méthode de segmentation doit prendre en compte l'ensemble des déformations possibles.
- L'apparence autour de la bouche peut être modifiée par la présence de moustaches ou de barbes (cf. Fig. 2.02.a), ou par la présence d'un objet (cf. Fig. 2.02.b).
- La bouche peut être partiellement (cf. Fig. 2.02.c) ou complètement (cf. Fig. 2.02.d) occultée par des moustaches ou par un objet.
- Les variations d'illumination peuvent affecter la détection en modifiant l'apparence autour et sur les lèvres. Une partie de la lèvre inférieure peut être surexposée ou plus brillante si la lumière vient de dessus (cf. 4ème image de la figure 2.01). La peau située juste en dessous de la bouche peut être également plus ou moins sombre suivant la direction de la lumière.
- Concernant le contour intérieur, nous verrons que la tâche est d'autant plus difficile que l'intérieur de la bouche peut être sombre (cavité orale), brillant (dent) ou d'une couleur proche de celle des lèvres (gencives et langue).

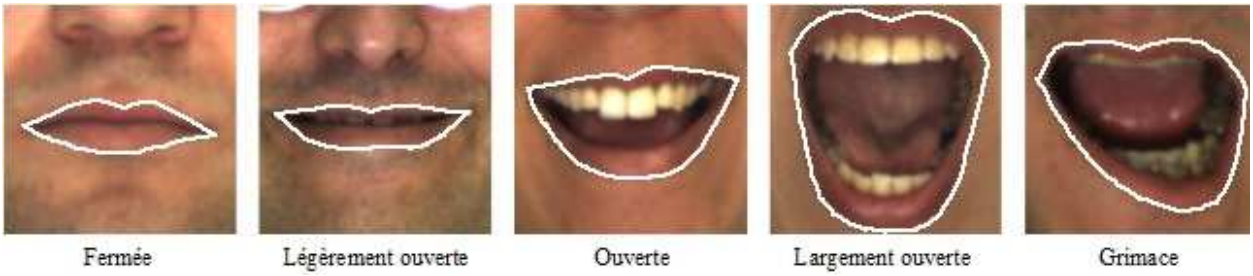


Fig. 2.01. Variations de la forme de la bouche [Martinez, 1998].

Fig. 2.02. Variations de l'apparence et occultations [Martinez, 1998].

Dans ce chapitre, nous faisons un état de l'art des méthodes de segmentation labiale que nous classons en deux catégories : les approches « région » et les approches « contour ».

Les méthodes de segmentation des lèvres utilisent essentiellement des informations de chrominance et la première étape est de choisir l'espace couleur le plus adapté, à savoir celui qui va permettre une bonne discrimination des pixels lèvre. Pour les approches région, l'espace couleur optimal doit permettre de représenter les pixels « lèvre » et les pixels « peau » en deux groupes homogènes bien séparés (faible variance intra-classe et forte variance inter-classe). La section 2.1 présente les espaces couleurs standards (*RGB, YCbCr, TLS*) utilisés fréquemment pour l'analyse labiale, mais également d'autres informations couleurs développées spécialement pour augmenter le contraste entre les lèvres et la

peau. En ce qui concerne les approches contour, il est nécessaire de choisir un espace couleur qui permette d'accentuer la frontière entre les lèvres et la peau. Les pixels appartenant au contour sont associés à une valeur forte du gradient. Dans la section 2.2, nous présentons différentes manières d'obtenir un fort gradient au niveau des contours labiaux.

Dans la section 2.3, nous détaillons plusieurs approches région que nous regroupons en trois catégories : les méthodes de seuillage, les méthodes de classification et les modèles statistiques de forme et d'apparence.

Les approches contour sont introduites dans la section 2.4. Cette famille de méthode utilise généralement deux types de modèles déformables : les contours actifs (ou snakes) et les modèles paramétriques. Dans cette section, les approches contour sont présentées brièvement et un état de l'art plus détaillé est proposé dans le chapitre 3 car ce sont des algorithmes de segmentation des lèvres orientés contour que nous avons développés.

Finalement, la section 2.5 est une conclusion sur les avantages et inconvénients des méthodes de segmentation des lèvres introduites dans ce chapitre.

2.1. Les espaces couleurs

Le choix d'un espace couleur adapté est une étape cruciale pour tout algorithme de segmentation. Dans le contexte de l'analyse labiale, l'espace couleur approprié est celui qui permet de distinguer les pixels lèvre des pixels peau. Dans cette partie, nous testons les performances de plusieurs espaces couleurs. En premier lieu, nous étudions les espaces standards *RGB, YCbCr* et *TLS*, qui s'avèrent in fine comme n'étant pas nécessairement les plus adaptés. Ensuite, nous présentons les pseudo-teintes qui sont des composantes couleurs développées spécialement pour accentuer la différence de contraste entre les pixels lèvre et les pixels peau.

Il est à noter que les résultats présentés proviennent d'une étude que nous avons menée avec Alice Caplier, Christian Bouvier et Pierre-Yves Coulon lors de l'écriture d'un chapitre de livre sur la segmentation des lèvres [Caplier, 2009]. Ils ont été obtenus suite à l'analyse d'échantillons de pixels lèvre et peau extraits sur une base de données, constituée de 150 images provenant de 20 sujets différents, et acquises avec la même caméra et les mêmes conditions d'illumination.

2.1.1. Les espaces couleurs standards

Les espaces couleurs standards les plus fréquemment employés en analyse labiale sont les espaces *RGB, YCbCr et TLS*. Dans notre étude, nous nous intéressons aux performances dans le contexte de la segmentation des lèvres. Pour une description plus détaillée de ces différents espaces couleurs, le lecteur pourra consulter, par exemple, les travaux de Ford *et al.* [Ford, 1998].

2.1.1.a. L'espace *RGB*

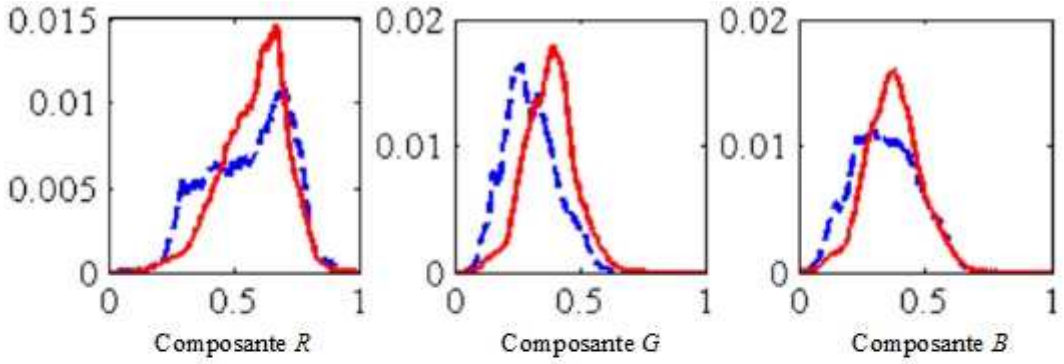
L'espace couleur *RGB* est un système de couleur additif basé sur la théorie trichromatique. En combinant les trois primitives Rouge, Verte et Bleue, il est possible d'obtenir presque toutes les couleurs visibles.

Plusieurs études ont proposé de travailler dans l'espace couleur *RGB* pour extraire des informations labiales ([Chan, 1998]; [Wojdel, 2001]; [Daubias, 2002]; [Nefian, 2002]; [Patterson, 2002]; [Chen, 2004]).

La figure 2.03.a montre l'histogramme des pixels lèvre et peau des échantillons de notre base de données pour les trois composantes *R, G* et *B*. Un exemple des trois composantes pour une image centrée sur la bouche est visible sur la figure 2.03.b.

Les composantes chromatiques *R, G* et *B* ont des distributions de valeurs larges aussi bien pour les pixels peau que pour les pixels lèvre. Pour chacune des trois composantes, les distributions associées aux lèvres et à la peau se chevauchent fortement. En conséquence, il est difficile de définir deux zones distinctes, l'une correspondante aux pixels peau et l'autre aux pixels lèvre. Même si les lèvres sont généralement vues plus rouges que la peau, les histogrammes de la figure 2.03.a montrent que la composante *R* est prédominante par rapport aux composantes *G* et *B*, mais qu'elle est aussi importante pour les lèvres que pour la peau. En outre, les pics des distributions de *R* et de *G* sont plus éloignés pour la peau que pour les lèvres, ce qui explique que la peau apparaît plus jaune que les lèvres. Nous verrons dans la section 2.2 que cette caractéristique sera exploitée pour créer des plans teintes spécifiques aux lèvres.

L'espace couleur *RGB* n'apparaît donc pas comme étant l'espace le plus approprié pour l'analyse labiale, d'autant plus que les informations de chrominance et de luminance sont mélangées, contrairement aux espaces *YCbCr* ou *TLS*.



a. Histogramme des pixels lèvres (ligne pointillée) et peau (ligne pleine) pour les 3 composantes *RGB*.

b. Exemple des 3 composantes *RGB* pour une même image de bouche.

Fig. 2.03. Espace couleur *RGB*.

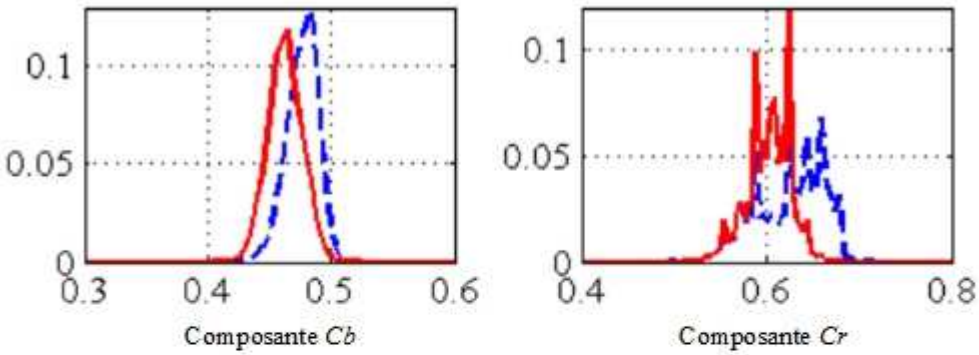
2.1.1.b Espace *YCbCr*

L'espace *YCbCr* est obtenu à partir de l'espace *RGB* par une transformation bijective. La luminance *Y* est séparée des composantes couleurs *Cb* et *Cr*. Une modification des valeurs des canaux *Cb* et *Cr* permet d'obtenir une couleur différente, tout en gardant la même luminance. Cet espace est utilisé notamment pour la télévision et les images JPEG.

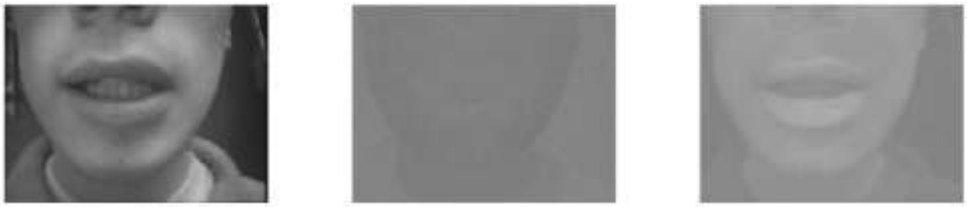
L'espace *YCbCr* est également beaucoup employé en analyse labiale ([Zhang, 1997]; [Tsapatsoulis, 2002]; [Gacon, 2005]).

La figure 2.04.a montre lhistogramme des pixels lèvre et peau issus de notre base de données pour les deux composantes de chrominance *Cb* et *Cr*. Un exemple des trois composantes de l'espace *YCbCr* pour une image centrée sur la bouche est visible sur la figure 2.04.b.

Les histogrammes montrent clairement un chevauchement des distributions associées aux lèvres et à la peau pour les composantes *Cb* et *Cr*. Sur la figure 2.04.b, ces deux composantes chromatiques sont peu efficaces pour séparer les pixels lèvre et les pixels peau. *Cb* apparaît même très bruitée et relativement uniforme sur l'ensemble du visage. L'espace couleur *YCbCr* n'est pas l'espace le plus adapté pour le problème de la segmentation des lèvres.



a. Histogramme des pixels lèvres (ligne pointillée) et peau (ligne pleine) pour les 2 composantes *Cb* et *Cr*.



b. Exemple des 3 composantes *YCbCr* pour une même image de bouche.

Fig. 2.04. Espace couleur *YCbCr*.

2.1.1.c. Espace *TLS*

L'espaces *TLS* (Teinte, Luminance, Saturation) est un espace de représentation des couleurs plus proche de la perception humaine. Plusieurs formules mathématiques ont défini des espaces différents comme *HSV* (*Hue, Saturation, Value*), *HSI* (*Hue, Saturation, Intensity*) ou *HSL* (*Hue, Saturation, Lightness*), mais ils décrivent trois informations similaires :

- la teinte : le type de couleur,
- la saturation : l'intensité de la couleur ou la pureté,
- la luminance : la brillance.

La teinte est sans doute la composante couleur la plus utilisée en analyse labiale ([Coianiz, 1996]; [Vogt, 1996]; [Chibelushi, 1997]; [Brand, 2001]; [Pantic, 2001]; [Yin, 2002]). Dans [Zhang, 2002], l'étude porte sur la comparaison des espaces *RGB, YCbCr* et *HSV* dans le contexte de la segmentation des lèvres. Les auteurs montrent que la composante teinte *H* permet une meilleure séparation des pixels lèvre et peau par rapport aux autres plans couleurs testés.

Dans les espaces *TLS*, la teinte est représentée par un angle dont les valeurs sont proches de la borne supérieure (2π) de l'intervalle de définition pour les objets à dominante rouge, tels que les lèvres. La figure 2.05.a montre l'histogramme de la composante *H*; les distributions sont décalées pour se situer au centre. Les distributions des pixels lèvre et peau sont relativement concentrées, mais elles se chevauchent beaucoup. Un exemple du plan teinte pour une image de bouche est visible sur la figure 2.05.a. On remarque qu'il est difficile de distinguer les lèvres par rapport au reste du visage.

2.1.2. Plans de couleur développés spécialement pour la segmentation des lèvres

Les espaces couleurs standards n'étant pas bien adaptés au problème de l'analyse labiale, plusieurs auteurs ont proposé des plans couleurs dédiés permettant d'accentuer la différence entre les pixels lèvre et les pixels peau. Ces plans sont calculés à partir des composantes de chrominance standards R et G.

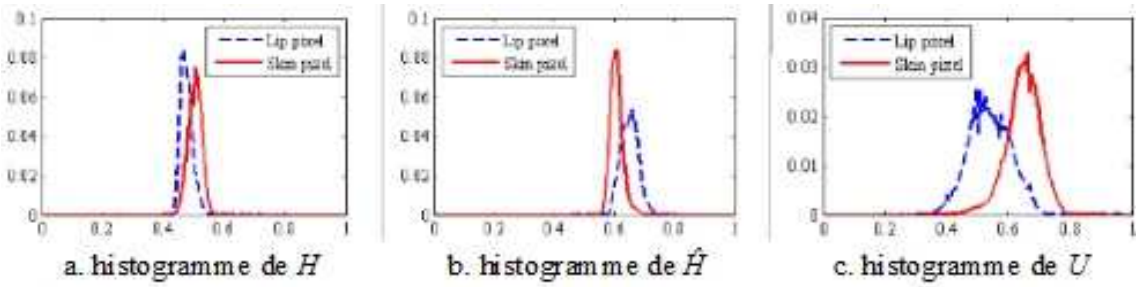


Fig. 2.05. Histogramme des pixels lèvre (ligne pointillée) et peau (ligne pleine) pour les composantes H, $\hat{H}$ et U.

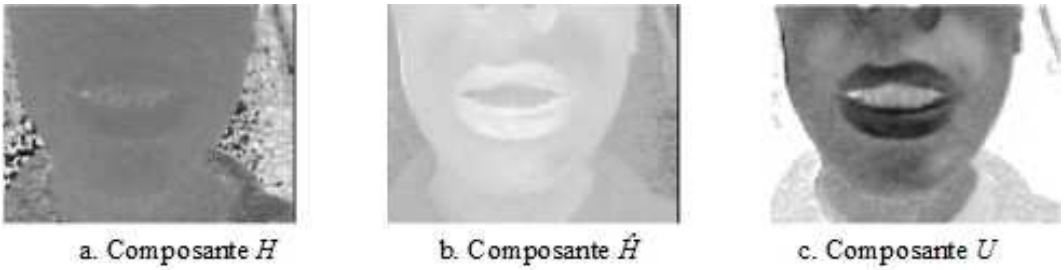


Fig. 2.06. Exemple des composantes H, $\hat{H}$ et U pour une même image de bouche.

2.1.2.a. La pseudo-teinte $\hat{H}$

Nous avons vu dans la section 2.1.1.a que, de manière générale, la différence entre les composantes R et G est plus grande pour les pixels lèvre que pour ceux de la peau.
Cette propriété est exploitée dans [Poggio, 1998] pour construire la pseudo-teinte $\hat{H}$ calculée de la manière suivante :

$$\hat{H} = \frac{R}{R+G} \tag{2.01}$$

Contrairement à la teinte classique H, la pseudo-teinte est bijective. La figure 2.05.b montre que, pour le cas de la pseudo-teinte, les distributions des valeurs des pixels lèvre et peau sont concentrées. Notre base de données fournit des résultats sensiblement équivalents à la teinte classique, mais le chevauchement est moins important. $\hat{H}$ prend une valeur plus forte pour les pixels lèvre que pour les pixels peau (cf. Fig. 2.06.b) et elle permet donc une bonne séparation des lèvres et de la peau. La pseudo-teinte est notamment utilisée dans les travaux de [Eveno, 2003] pour extraire le contour extérieur des lèvres.

2.1.2.b. Le canal U

Dans [Liévin, 2004], les auteurs proposent un plan teinte calculé de la manière suivante :

$$U = \begin{cases} 256x\frac{R}{G} & \text{si } R > G \\ 255 & \text{sinon} \end{cases} \quad \text{où } R \text{ et } G \in [0, 256[\tag{2.02}$$

Les figures 2.05.c et 2.06.c montrent que la teinte *U* permet une bonne séparation des pixels lèvre et peau. *U* est utilisée dans les travaux ([Liévin, 2004]; [Beaumesnil, 2006]; [Bouvier, 2007]) pour la segmentation des contours des lèvres.
Dans [Chiou, 1997], les auteurs augmentent le contraste entre la bouche et la peau avec le rapport *G*/*R*. Les performances de discrimination sont équivalentes à la composante *U*, mais aucune condition de restriction (condition *R*>*G*, cf. Eq. 2.02) n'est imposée.

2.2. Accentuation du contour des lèvres : calcul de gradient

Lorsque l'on désire extraire les contours des lèvres, il est nécessaire d'accentuer la frontière entre les lèvres et la peau. Pour obtenir un fort gradient sur les contours des lèvres, il faut que l'espace de représentation de l'image permette une variation importante des valeurs des pixels lèvres et peau. Ainsi plusieurs travaux utilisent les espaces couleurs introduits dans la section 2.1.

2.2.1. Gradient intensité

Du fait de la limitation des puissances de calcul et des techniques de traitement d'images de l'époque, les premiers travaux traitant du problème de la détection des contours des lèvres utilisaient essentiellement le gradient intensité ([Hennecke, 1994]; [Radeva, 1995]). Mais même des études plus récentes ([Pardas, 2001]; [Delmas, 2002]; [Seyedarabi, 2006]; [Werda, 2007]) calculent le gradient à partir du plan intensité car la région de la bouche est caractérisée par des changement d'illumination entre les lèvres et la peau. Par exemple, dans les conditions d'éclairage les plus courantes, la source de lumière vient d'en haut et la frontière supérieure de la bouche est un contour avec une forte luminance alors que la lèvre supérieure est plus sombre. De la même manière, la frontière inférieure de la bouche est un contour avec une faible luminance alors que la lèvre inférieure est bien éclairée.

En outre, le gradient intensité est une information particulièrement adaptée à la segmentation du contour labial intérieur car l'intérieur de la bouche contient des zones très claires (dents) et très sombres (cavité orale).

Suivant le signe du gradient, le contour accentué est soit la frontière d'une zone sombre au dessus d'une zone brillante, soit la frontière d'une zone sombre au dessous d'une zone brillante. La figure 2.07 montre un exemple d'accentuation de contours au niveau de la bouche en utilisant le plan luminance (les frontières avec un fort gradient sont en blanc).

Fig. 2.07. Gradient intensité.

2.2.2. Gradient couleur

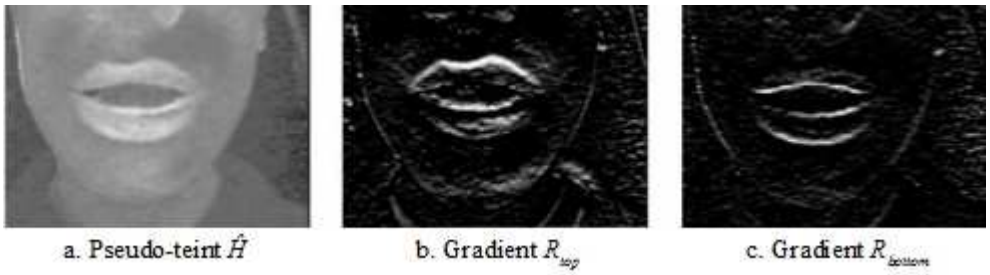
Les espaces couleurs standards sont peu utilisés pour calculer des gradients directement à partir des composantes couleurs; nous avons vu dans la section 2.1 qu'ils n'étaient pas particulièrement adaptés à l'analyse labiale. La solution la plus envisagée consiste à calculer le gradient à partir des plans teinte développés pour accentuer le contraste entre les lèvres et la peau (cf. section 2.1.2).

Dans [Eveno, 2003], deux gradients couleurs sont construits à partir de la pseudo-teinte $\hat{H}$ (cf. Eq. 2.01) pour extraire le contour extérieur des lèvres. R_{top} permet d'accentuer le contour extérieur supérieur et R_{bottom}, le contour inférieur (cf. Figure 2.08) :

$$\begin{aligned} R_{top} &= \nabla(\hat{H} - L) \\ R_{bottom} &= \nabla(\hat{H}) \end{aligned} \tag{2.03}$$

où $\hat{H}$ est la pseudo-teinte et L est le plan luminance. ∇ est l'opérateur gradient.

$\hat{H}$ est utilisée pour son pouvoir discriminant (cf. Fig. 2.08.a). Pour le contour supérieur, la pseudo-teinte est combinée avec la luminance en considérant que la lumière vient d'en haut et que la frontière supérieure est une zone de forte luminance. Beaumesnil *et al.* [Beaumesnil, 2006] utilisent le même genre de combinaison pour le contour extérieur des lèvres, mais le plan teinte U (cf. Eq. 2.02) remplace la pseudo-teinte.



a. Pseudo-teint $\hat{H}$ b. Gradient R_{top} c. Gradient R_{bottom}

Fig. 2.08. Gradient couleur [Eveno, 2003].

2.2.3. Gradient calculé à partir d'une carte de probabilité

Dans [Vogt, 1996] et [Liévin, 2004], le gradient est obtenu à partir d'une carte de probabilité, qui exprime la probabilité qu'un pixel appartienne aux lèvres. Ceci permet de séparer les pixels lèvres des pixels peau.
Dans [Vogt, 1996], la carte de probabilité est construite à partir des composantes H et S de l'espace couleur *HIS*. L'image obtenue donne des valeurs élevées pour les pixels ayant une forte probabilité d'appartenir aux lèvres.
Dans [Liévin, 2004], la carte de probabilité est obtenue à l'aide d'un modèle de champ de Markov aléatoire (**M**arkov **R**andom **F**ield) combinant le plan teinte U (cf. Eq. 2.02) et une information de mouvement. Le modèle MRF fournit une image en niveau de gris regroupant les régions du visage en différents clusters, dont un cluster particulier pour les pixels de la bouche.

2.2.4. Gradient calculé à partir d'une image binaire

Une autre possibilité consiste à calculer le gradient à partir d'une image binaire où les lèvres sont représentées par des pixels blancs et le reste du visage par des pixels noirs.
Ce type de méthode est utilisé dans [Wark, 1998] et [Yokogawa, 2007], mais le résultat dépend de la précision de la construction de l'image binaire.

2.3. Méthodes de segmentation des lèvres basées région

Les méthodes de segmentation des lèvres par une approche basée région peuvent se classer en trois catégories principales : les méthodes de seuillage, les méthodes de classification et les modèles statistiques.

2.3.1. Les méthodes de seuillage

Les méthodes de seuillage sont des approches bas niveau qui exploitent des informations colorimétriques ou de luminance en n'utilisant aucune information sur la forme ou les contours des lèvres. Ce type d'approche considère que la bouche est représentée par un groupe de pixels spécifique et homogène dans un espace couleur donné.

Dans [Petajan, 1984], Petajan localise la position des narines et de la région de la bouche à l'aide de mesures morphologiques. Un simple seuillage du plan luminance permet de segmenter les lèvres et de mesurer des grandeurs caractéristiques (hauteur, largeur, surface) pour une application de reconnaissance automatique de la parole. La même idée est utilisée dans [Lyons, 2003] en ajoutant un deuxième seuillage sur la composante *R*, car selon Lyons, l'intérieur de la bouche est rouge et sombre. Cette approche fournit des informations sur l'ouverture de la bouche. Coianiz *et al.* [Coianiz, 1996] segmentent les lèvres en seuillant le plan teinte *H*. Dans [Zhang, 2000], les auteurs obtiennent un masque binaire de la bouche à l'aide de deux seuillages effectués sur *H* et *S* (cf. Fig. 2.09).

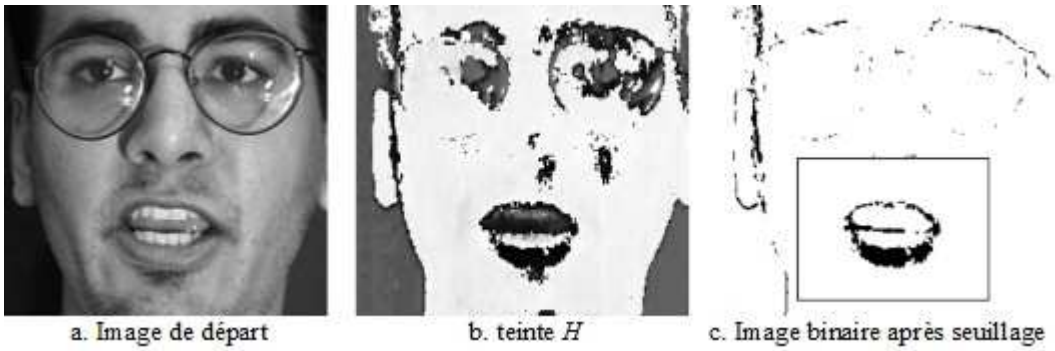


Fig. 2.09. Seuillage effectué dans [Zhang, 2000].

Un seuillage haut et un seuillage bas sur le rapport R/G (cf. section 2.1.2) sont utilisés dans [Chiou, 1997]. Dans [Wark, 1998], ce même seuillage est complété par des opérations morphologiques pour obtenir une segmentation moins bruitée (cf. Fig. 2.10). R/G est également employé dans [Lucey, 1999].

Fig. 2.10. Seuillage effectué dans [Wark, 1998].

Ce genre d'approche est peu fiable car les seuils sont difficiles à fixer automatiquement. De manière générale, les seuils sont choisis par un expert humain pour une base d'images donnée, mais les mêmes seuils sont rarement réutilisables lorsque les conditions d'acquisition ou les sujets changent. Un simple seuillage peut servir à calculer des paramètres labiaux, mais il ne permet pas d'obtenir des contours précis, car la région segmentée obtenue est souvent bruitée. Certains auteurs ont proposé de maquiller les lèvres en bleu. Avec un tel artifice, un seuillage sur le plan teinte donne des résultats très satisfaisants. Même si cette astuce est difficilement utilisable dans la pratique, ceci a permis de mettre en évidence le potentiel biométrique de l'analyse labiale ([Chibelushi, 1997]; [Brand, 2001]).

En conclusion, les techniques de seuillage sont souvent utilisées comme première étape en vue de détecter la zone d'intérêt ou pour ensuite lisser les contours avec un modèle de forme (un modèle paramétrique par exemple, cf. section 2.4.2).

2.3.2. Les méthodes de classification

Les méthodes de classification permettent de découper une image en plusieurs groupes homogènes ou classes. Les premières étapes consistent à définir les classes, les caractéristiques de chaque classe et la méthode de classification. Dans le contexte de la segmentation des lèvres, cela revient souvent à définir deux groupes de pixels appartenant aux lèvres ou à la peau. Nous pouvons distinguer deux approches différentes : les méthodes supervisées et les méthodes non supervisées.

2.3.2.a. Classification supervisée

Les approches supervisées nécessitent des connaissances *a priori* sur les classes à segmenter. Les caractéristiques sont obtenues à l'aide d'un apprentissage. La conception de la base d'apprentissage est une étape essentielle et elle doit couvrir un large éventail de possibilités (différentes conditions d'éclairage, résolutions...) pour obtenir un algorithme robuste. Les méthodes de classification rencontrées dans la littérature sont essentiellement basées sur des approches statistiques, des réseaux de neurones ou des machines SVM.

Dans [Chan, 1998], les auteurs utilisent une combinaison linéaire des composantes *R*, *G* et *B* pour accentuer le contraste entre les lèvres et la peau. Les paramètres du modèle linéaire sont obtenus par une analyse statistique d'échantillons de pixels lèvre et de pixels peau. Finalement, un seuillage est effectué sur l'image obtenue par la combinaison des trois plans pour extraire les lèvres. La même étude est proposée dans [Nefian, 2002]. Les paramètres sont définis par une analyse discriminante linéaire et un masque binaire de la bouche est construit par seuillage des plans combinés (cf. Fig. 2.11).

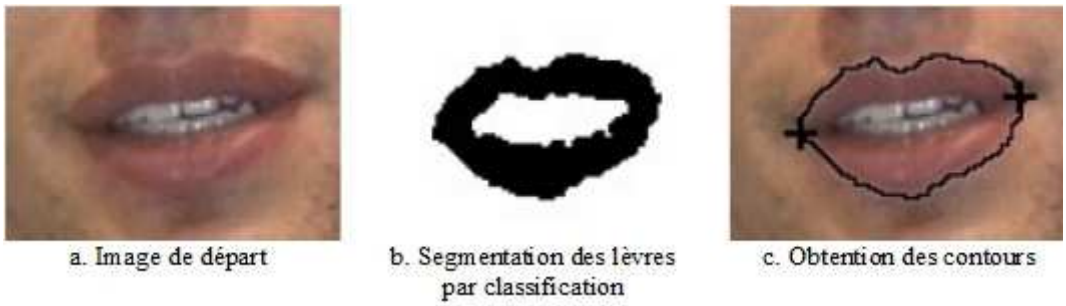


Fig. 2.11. Extraction des lèvres par classification supervisée [Nefian, 2002].

Patterson *et al.* [Patterson, 2002] utilisent une base d'apprentissage de lèvres et de visages (cf. Fig. 2.12.a) pour déterminer des approximations gaussiennes des distributions des pixels lèvre, visage et fond dans l'espace *RGB*. Finalement, la classification des pixels d'une image inconnue dans l'une des trois classes est réalisée avec un classifieur de Bayes (cf. Fig. 2.12). Une approche similaire est suivie dans

[Gacon, 2005]. Gacon *et al.* construisent un modèle colorimétrique basé sur un mélange de gaussiennes pour discriminer les pixels en trois classes : lèvres, peau et non affecté. Dans cette étude, l'information chromatique utilisée est la pseudo-teinte $\hat{H}$ (cf. Eq. 2.01).

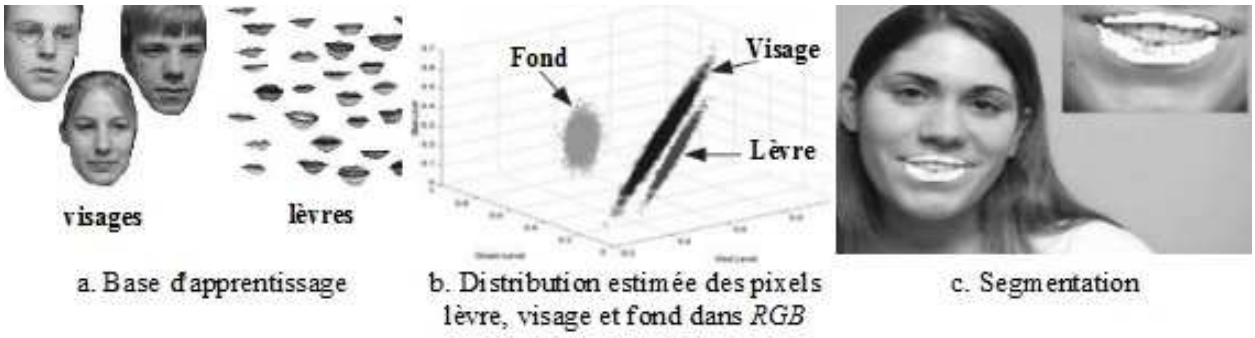


Fig. 2.12. Classification par approximations gaussiennes [Patterson, 2002].

Dans [Wojdel, 2001], un réseau de neurones est utilisé pour déterminer si un pixel appartient ou non aux lèvres. Le réseau est composé de trois couches et il est entraîné avec une base d'apprentissage de lèvres segmentées manuellement. Les pixels de l'image en *RGB* sont donnés en entrée, et le réseau de neurones fournit 0 ou 1 en sortie selon que le pixel est un pixel lèvre ou non. Daubias *et al.* [Daubias, 2002] entraîne également un réseau de neurones multi-couches pour classer des pixels *RGB* en trois classes : lèvre, peau et intérieur de la bouche.
Dans [Castañeda, 2005], les auteurs utilisent un SVM pour détecter le visage et les positions de caractéristiques faciales telles que les lèvres et les yeux.

Dans [Yang, 1996], Yang *et al.* montrent que, même avec des conditions d'éclairage constantes, un changement de caméra entraîne des variations du rendu des couleurs. Ceci confirme la difficulté du choix de la base d'apprentissage qui doit prendre en compte les possibles variations d'éclairage, de caméra, d'échelle et de sujet. En outre, les données d'apprentissage sont généralement obtenues manuellement; l'étiquetage devient vite lourd et fastidieux. D'autres études préfèrent s'orienter vers une classification non supervisée.

2.3.2.b. Classification non supervisée

Les approches non supervisées ne nécessitent aucune connaissance *a priori* sur les distributions statistiques. En ce qui concerne l'analyse labiale, ce type de méthode considère que même si les lèvres et la peau n'ont pas une couleur constante pour chaque individu, leur différence est telle qu'il est possible de séparer les pixels lèvre et les pixels peau sans avoir effectué d'apprentissage au préalable.

Une approche combinant des informations de couleur et de gradient est proposée dans [Bouvier, 2007] pour réaliser une segmentation non supervisée des lèvres. Un modèle de mélange de gaussiennes est construit en utilisant le plan teinte U pour décrire la distribution colorimétrique de la zone de la bouche (cf. Eq. 2.02). Les images étant centrées sur le bas du visage, il y a plus de pixels peau que de pixels lèvres, et la gaussienne avec le poids le plus élevé est associée à la peau. Le modèle permet d'établir une carte d'appartenance des pixels de l'image aux lèvres. Plus un pixel a des chances d'appartenir aux lèvres, plus il est blanc sur la carte (cf. Fig. 2.13.b). Finalement, un seuillage utilisant les informations de gradient R_{top} et R_{bottom} (cf. Eq. 2.03) est réalisé pour obtenir un masque binaire des lèvres (cf. Fig. 2.13.c). Un modèle de mélange de gaussiennes est également étudié dans [Tian, 2000a] pour analyser la distribution colorimétrique des lèvres.

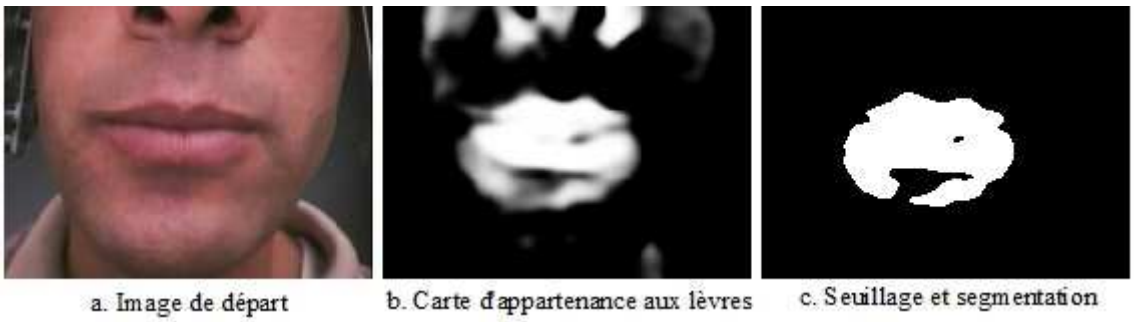


Fig. 2.13. Classification non supervisée par mélange de gaussiennes [Bouvier, 2007].

Dans [Liévin, 2004], une classification non supervisée est basée sur les champs de Markov aléatoires (MRF). Cette étude combine une information de couleur, représentée par la teinte *U*, et une information de mouvement. L'algorithme est itératif et commence par séparer une classe en calculant la teinte *U* de l'image; l'information de mouvement est obtenue par la différence $|I_t-I_{t-1}|$, où I_t est l'image courante et I_{t-1} l'image à l'instant précédent. La classification est réalisée en associant les deux informations. S'ils restent des pixels à classer, une seconde itération est appliquée. Dans le cadre de la segmentation des lèvres, ce travail consiste à séparer 2 classes : les lèvres et la peau, et le processus nécessite donc deux itérations. La première itération concerne les pixels peau, sachant que l'image est centrée sur le bas du visage et qu'il y a plus de pixels peau que de pixels lèvre. La figure 2.14 montre des exemples de segmentation obtenue à la deuxième itération.

Fig. 2.14. Classification non supervisée par champ de Markov aléatoire [Liévin, 2004].

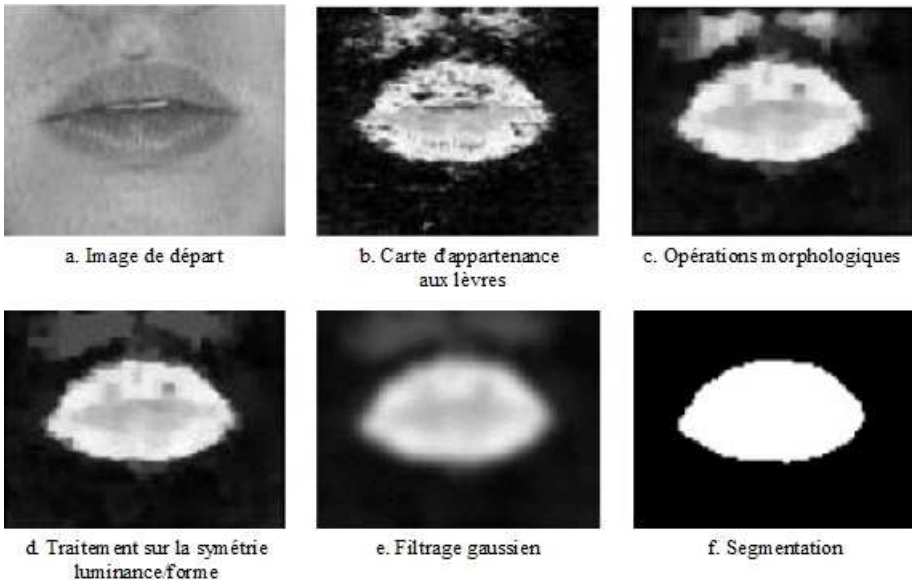


Fig. 2.15. Classification non supervisée avec un algorithme Fuzzy C-mean [Liew, 2003].

Une séparation des pixels lèvre et des pixels peau est effectuée avec un algorithme Fuzzy C-mean par Liew *et al.* [Liew, 2003]. Un vecteur de chrominance basé sur les espaces couleurs *CIELuv* et *CIELab* [Wyszecki, 1982] est associé à chaque pixel de l'image et une carte d'appartenance aux lèvres est construite (cf. Fig. 2.15.b). Ensuite, une série de traitements (opérations morphologiques, contraintes symétriques sur la forme et la luminance, et filtrage gaussien) permet de lisser le résultat et d'éviter des erreurs de classification. Un seuillage sur les degrés d'appartenance donne un masque binaire des lèvres (cf. Fig. 2.15.f). Dans [Leung, 2004] et [Wang, 2007], les Fuzzy C-means sont également employés. Une contrainte de forme est ajoutée en supposant que les lèvres ont une allure elliptique.

Dans cette section, nous avons présenté les méthodes de classification. La classification peut être supervisée (connaissances *a priori* sur les classes à segmenter) ou non supervisée (aucune connaissance préalable sur les classes à segmenter). Les méthodes de classification ne permettent pas une segmentation précise et ce type d'approche est souvent utilisé pour obtenir une boîte de la bouche ou définir des paramètres labiaux.

2.3.3. Les modèles statistiques

Les modèles statistiques sont des méthodes supervisées et ils sont construits avec une base d'apprentissage. Contrairement aux méthodes de classification présentées précédemment, les modèles statistiques sont entraînés pour décrire la forme et l'apparence des lèvres, et non leur distribution colorimétrique. Les **M**odèles **A**ctifs de **F**orme (ASM) ont été proposés en premier par Cootes *et al.* [Cootes, 1995] pour segmenter des objets dans des images médicales en utilisant des connaissances *a priori* sur les formes admissibles. Les **M**odèles **A**ctifs d'**A**pparence (AAM) modélisent l'apparence de l'objet à extraire et ils proviennent également des travaux de Cootes [Cootes, 1998].

2.3.3.a. Les modèles statistiques de forme

Les ASM [Cootes, 1995] sont une application des modèles de distribution de points (**P**oint **D**istribution **M**odel, PDM, [Cootes, 1992]) pour extraire un objet d'une image. Les PDM modélisent les contours de l'objet à partir de plusieurs exemples d'apprentissage. Les M images d'apprentissage représentent les formes possibles de l'objet. Pour chaque image i d'apprentissage, N points sont localisés manuellement sur le contour de l'objet et un vecteur x_i contenant les coordonnées des N points est disponible. On dispose donc de M vecteurs x_i qui, après normalisation, constituent la base d'apprentissage. Une **A**nalyse en **C**omposantes **P**rincipales (ACP) est effectuée sur la base d'apprentissage pour calculer les modes de variation. A partir de la forme moyenne de l'objet obtenue avec :

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i \tag{2.04}$$

chaque forme x de l'objet peut être représentée de la manière suivante :

$$x = \bar{x} + P_s b_s \tag{2.05}$$

où $P_s=[p_1, p_2, \ldots, p_n]$ est la matrice des n modes de variation les plus significatifs, et $b_s=[b_1, b_2, \ldots, b_n]$ est un vecteur contenant les poids affectés à chaque mode propre. On considère que les poids b_i ont une distribution gaussienne et qu'ils varient entre $-3\sigma_i$ et $+3\sigma_i$, où σ_i est l'écart type associé au vecteur propre p_i. En général, les n vecteurs propres permettent de garder 95% de la variance totale des données. L'ACP permet une forte réduction des dimensions et une forme est définie par un faible jeu de paramètres.

La convergence d'un modèle ASM est réalisée de manière itérative en minimisant ou maximisant une fonction de coût qui peut être basée sur un critère de gradient, de niveau de gris ou de flux optique par exemple.

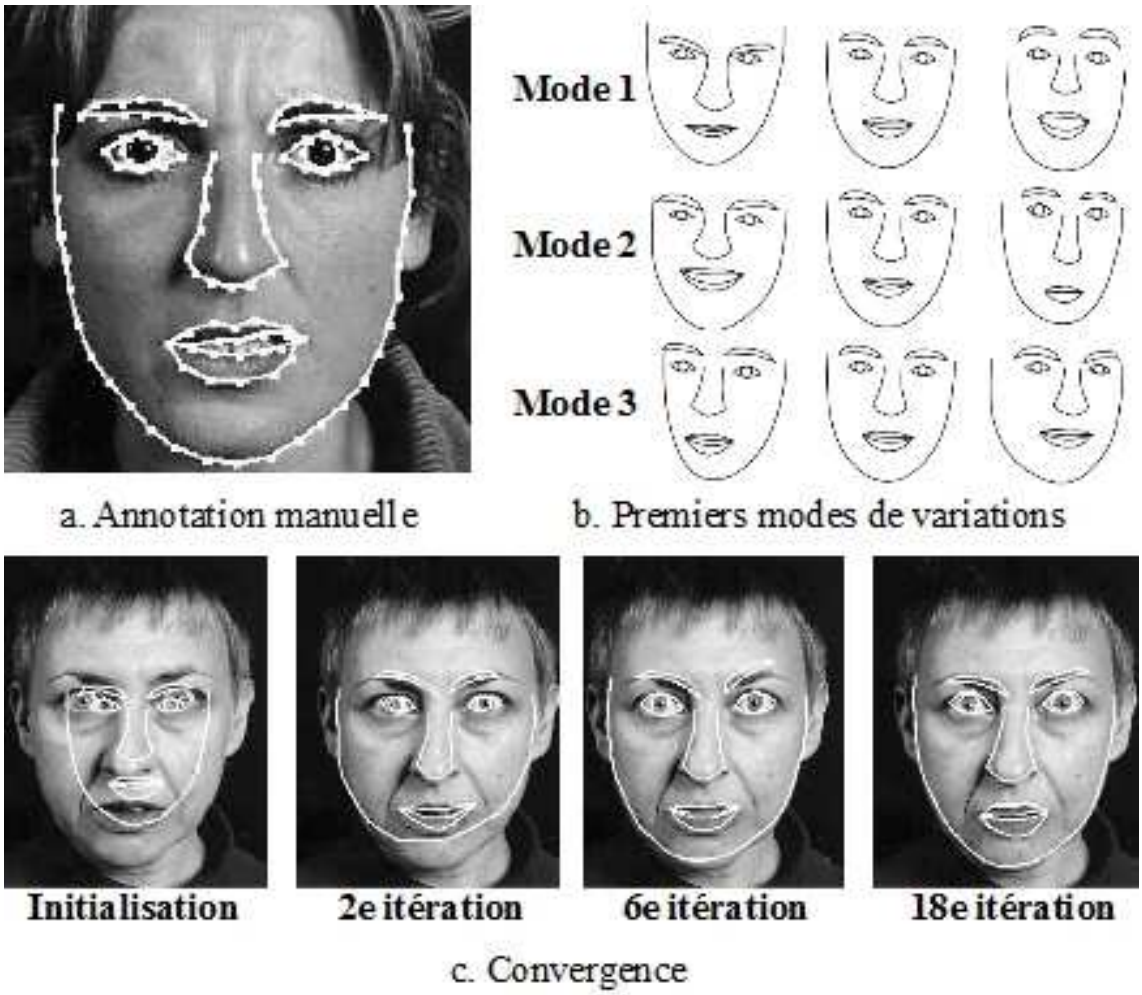
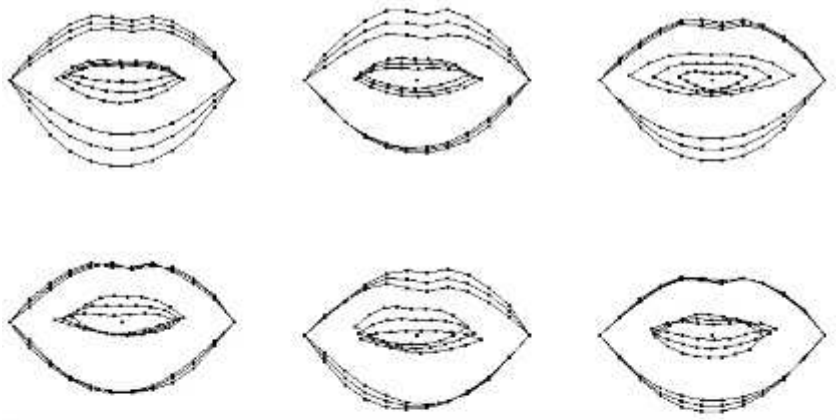


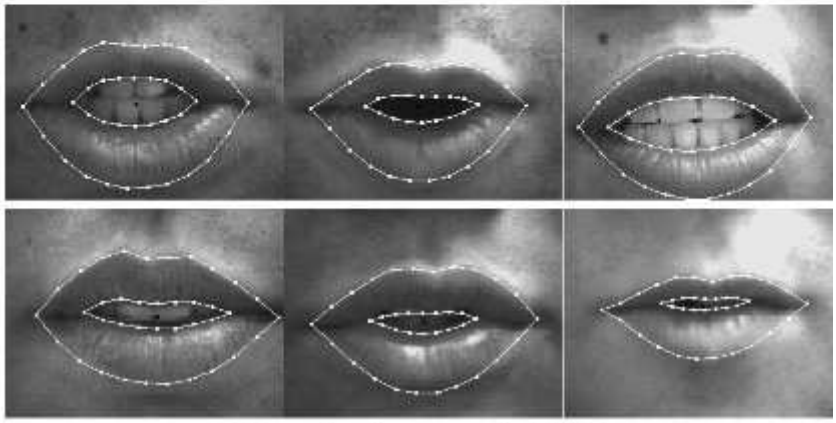
Fig. 2.16. Modèle de forme ASM pour l'analyse faciale [Cootes, 2004].

Dans [Cootes, 2004], un ASM est construit pour modéliser des traits permanents du visage (contour du visage, des yeux, des sourcils, du nez et des lèvres). Les modes de variation sont obtenus à partir d'images étiquetées manuellement (cf. Fig. 2.16.a et 2.16.b). La convergence est réalisée en maximisant un flux de gradient à travers une courbe (cf. Fig. 2.16.c).

Les ASM sont utilisés dans de nombreuses études pour détecter les formes des lèvres. Luettin *et al.* [Luettin, 1996] détectent les contours extérieur et intérieur des lèvres en utilisant un ASM pour une application d'identification de personne. Les six premiers modes de variation obtenus avec une ACP sont représentés sur la figure 2.17.a. Dans [Li, 2006], un algorithme ASM permet l'extraction des contours de la bouche. L'initialisation est réalisée avec une méthode de prédiction de la forme basée sur des contraintes de texture. Les informations de texture autour de chaque point du modèle sont caractérisées par un classifieur AdaBoost. Jang *et al.* [Jang, 2006] exploitent le même type d'approche pour segmenter les contours extérieur et intérieur des lèvres. La différence vient de l'utilisation d'un modèle de mélange de gaussiennes pour modéliser la texture autour des points de référence.



a. Les 6 premiers modes de variations



b. Exemples de segmentation labiale

Fig. 2.17. Modèle de forme ASM pour l'analyse labiale [Luettin, 1996].

2.3.3.b. Les modèles statistiques d'apparence

En plus de la forme, Cootes [Cootes, 1998] propose de modéliser également l'apparence. L'objectif est de combiner deux modèles statistiques actifs décrivant la forme et les niveaux de gris. Les AAM reposent sur les mêmes bases théoriques que les ASM et ils utilisent également une base d'apprentissage.

Un ASM est défini à partir de la base d'apprentissage. Pour construire le modèle statistique des niveau de gris, les images d'apprentissage sont déformées par une méthode de triangulation pour faire coïncider les points du modèle ASM avec ceux de la forme moyenne. Les niveaux de gris des pixels localisés à l'intérieur des contours de l'objet sont échantillonnés et une ACP donne les modes de variation de la luminance. Finalement, un AAM est construit par la concaténation des paramètres de forme et de luminance des images de la base. La convergence des AAM s'effectue aussi de manière itérative en minimisant la distance entre l'apparence générée et l'image à traiter.

La figure 2.18 illustre les trois premiers modes de variation du modèle AAM employé par Luettin [Luettin, 1997] pour la segmentation des lèvres. On remarque que les modes représentent des changements de forme et d'apparence. Nous pouvons également mentionner les travaux ([Matthews, 1998]; [Matthews, 2003]; [Gacon, 2005]) qui utilisent les AAM pour de l'analyse labiale.

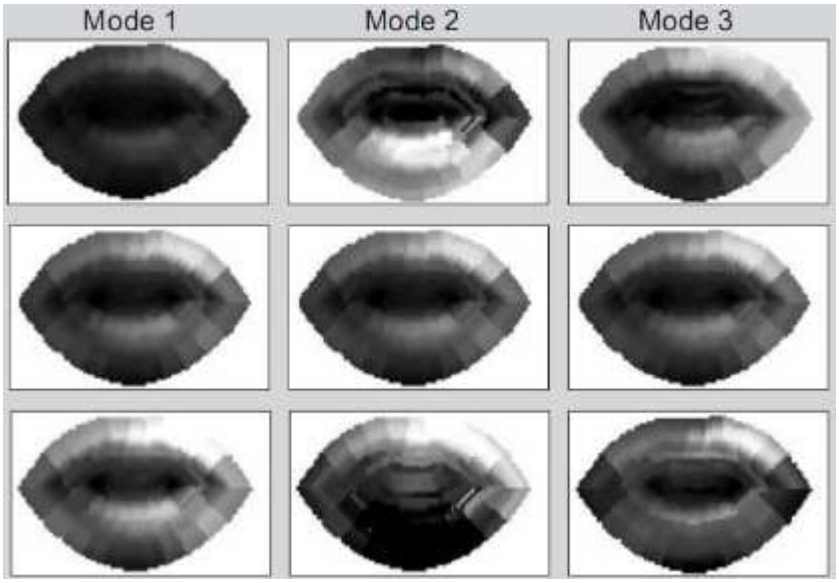


Fig. 2.18. Les 3 premiers modes de variation de l'apparence [Luettin, 1997].

Les ASM et AAM sont des méthodes qui permettent de fournir une segmentation toujours admissible car ils utilisent des connaissances obtenues par apprentissage sur la forme et l'apparence des lèvres. Cependant, la base d'entraînement doit être suffisamment importante pour couvrir tous les cas possibles. La liberté des déformations est limitée par ce que le modèle a appris. Et chaque cas doit être représenté assez souvent dans la base pour qu'il ne soit pas considéré comme secondaire. En outre, l'étiquetage manuel des images d'apprentissage est souvent lourd et fastidieux, dans la mesure où il faut positionner plusieurs dizaines de points sur plusieurs centaines d'images. Le résultat de la segmentation est également dépendant des conditions d'acquisition et les algorithmes basés sur des ASM et AAM sont peu robustes à des changements d'illumination. Cependant, la sensibilité des AAM aux conditions d'illumination peut être améliorée avec des pré-traitements, comme l'utilisation de cartes de distance [Le Gallou, 2006].

Dans cette section, nous avons présenté les modèles statistiques (ASM et AAM). Ces méthodes permettent une plus grande précision que les techniques de classification (cf. section 2.3.2), mais le résultat de la segmentation dépend de la base d'apprentissage qui doit prendre en compte tous les cas possibles au cours de l'entraînement. De plus, les contours étant représentés par un ensemble de points plus ou moins espacés, le résultat ne semble généralement pas naturel.

2.4. Méthodes de segmentation des lèvres basées contour

Les méthodes de segmentation orientées contour sont principalement basées sur les modèles déformables. Un modèle déformable consiste à élaborer et placer un modèle mathématique approprié du contour (une spline ou une courbe paramétrique par exemple) dans l'espace des données et à le déformer pour qu'il se bloque sur les frontières de l'objet d'intérêt. La déformation du contour initial est obtenue en minimisant un modèle d'énergie composé d'un terme interne, relié aux propriétés géométriques de la courbe, et un terme externe, calculé à partir des données de l'image. Nous pouvons séparer les modèles déformables en deux catégories : les contours actifs et les modèles paramétriques. Les contours actifs détectent les contours à partir d'un contour chaîné initial en modifiant un par un la position des points du modèle. Ce sont des algorithmes à forme libre, c'est-à-dire qu'aucune information *a priori* sur la forme de l'objet n'est prise en compte. Les modèles paramétriques sont définis par une description paramétrée de l'objet en intégrant des connaissances *a priori* sur la forme globale, ce qui permet d'obtenir un résultat admissible en modifiant les paramètres du modèle.

2.4.1. Les contours actifs ou snake : brève introduction

Les contours actifs, introduits par Kass *et al.* [Kass, 1987], sont communément appelés snake à cause de leur manière d'onduler comme un serpent durant leur déformation. Les snakes sont composés d'une série de points mobiles placés sur une courbe 2D. Selon l'application, la courbe peut être fermée ou non, avec des extrémités fixes ou non. Les contours actifs évoluent de manière itérative d'une position initiale jusqu'à leur position finale, en étant attirés par le minimum local le plus proche de la fonctionnelle d'énergie. La méthode de minimisation est commandée par des contraintes et elle est contrôlée par les données saillantes de l'image.

En ce qui concerne l'analyse labiale, les contours actifs sont largement utilisés car ils proposent d'importantes propriétés de déformation et un résultat de segmentation réaliste. Ils permettent également une implémentation facile et une convergence rapide, ce qui les rendent particulièrement efficace pour des applications de suivi des contours des lèvres.

Un snake est représenté par une courbe paramétrée *v* (*v(s)=(x(s), y(s)*, où *s* est l'abscisse curviligne) et une fonctionnelle d'énergie définie de la manière suivante :

$$\varphi(v)=\int \left(E_{\text{int}}(v(s))+E_{ext}(v(s))\right)ds \tag{2.06}$$

L'énergie interne E_{int} correspond aux propriétés mécaniques de la courbe et elle permet une certaine régularisation pendant les déformations. Les forces internes du snake imposent des contraintes sur la forme globale du modèle après convergence, en agissant localement autour de chaque point. L'énergie externe E_{ext} est liée aux données et elle déforme la courbe en fonction des caractéristiques saillantes de l'image, telles que les contours.

- Les contraintes internes doivent être suffisamment fortes du fait de la grande flexibilité des déformations de ce type de modèle à forme libre. L'énergie interne est définie comme étant :

$$E_{\text{int}}(v)=\frac{1}{2}\int \left(\alpha(s)|v'(s)|^2+\beta(s)|v''(s)|^2\right)ds \tag{2.07}$$

  où *v'(s)* et *v''(s)* sont les dérivées première et seconde de *v(s)*. $\alpha(s)$ est le coefficient d'élasticité et il représente la tension. Il intervient sur la longueur de la courbe. Une forte valeur de α impose de fortes contraintes de tension, alors que la courbe peut avoir des discontinuités dans le cas où α=0. $\beta(s)$ est le coefficient de rigidité et il représente la courbure. Lorsque β=0, la courbe peut présenter des parties convexes, alors qu'elle est plus lisse quand β est important.

- L'énergie externe du snake est calculée à partir des caractéristiques de l'image. De manière générale, l'énergie externe est fonction d'une information de gradient (pour attirer le snake vers les contours) ou du plan intensité (pour attirer le snake vers des zones claires ou sombres).

La courbe initiale se déforme de manière itérative en minimisant la fonction d'énergie. Ceci est un problème d'optimisation d'une fonctionnelle de plusieurs variables, qui, dans ce cas, sont les points du snake.

Les contours actifs sont largement utilisés pour des applications de détection de contours, du fait de leur capacité à intégrer les deux étapes d'extraction et de chaînage en une seule opération. Les snakes peuvent être utilisés aussi bien pour des contours ouverts que pour des contours fermés ou des contours avec des extrémités fixes. Les contours actifs sont rapides et simples à implémenter en 2D. Ils transforment un problème complexe de minimisation en une itération d'un système de matrice linéaire.

Un avantage supplémentaire est leur stabilité numérique face aux contraintes internes.

Cependant, l'utilisation des contours actifs présente quelques inconvénients. L'initialisation est la principale difficulté et elle peut amener facilement à une mauvaise segmentation. Les snakes étant des algorithmes à forme libre, ils sont attirés aveuglément vers le plus proche minimum local de la fonction d'énergie. En outre, les paramètres des contours actifs sont difficiles à régler et ils sont généralement choisis de manière heuristique. On peut également citer trois problèmes supplémentaires : la difficulté de converger vers des frontières concaves, l'instabilité face aux contraintes externes (les snakes peuvent dépasser le contour si les paramètres sont mal réglés) et le fait que le changement de topologie d'un objet ne peut être pris en compte. Plusieurs travaux ont proposé d'améliorer les performances des contours actifs standards introduits par Kass *et al.* On peut mentionner les forces ballon [Cohen, 1991], le flux du vecteur du gradient (**G**radient **V**ector **F**low, GVF, [Xu, 1998]) et les snakes génétiques [Ballerini, 1999].

Les contours actifs ont été étudiés de manière intensive dans le cadre de l'analyse labiale ([Shinchi, 1998]; [Delmas, 2002]; [Seguier, 2003]; [Kuo, 2005]; [Beaumesnil, 2006]; [Seyedarabi, 2006]...). Ces travaux seront détaillés dans le chapitre 3.

2.4.2. Les modèles paramétriques : brève introduction

Les modèles paramétriques possèdent plusieurs similarités avec les contours actifs. Ce type de modèle déformable évolue vers les contours de l'objet d'intérêt avec un algorithme de minimisation d'énergie. L'énergie est également composée d'un terme interne et d'un terme externe. La principale différence est l'utilisation de connaissances *a priori* sur la forme globale de l'objet à extraire.

Les modèles paramétriques, introduits par Yuille *et al.* [Yuille, 1992], décrivent une forme avec un modèle paramétré, composé généralement d'un assortiment de courbes (cercles, ellipses, paraboles, courbes cubiques...). Le modèle interagit dynamiquement avec les données à travers une fonction de coût. La convergence est obtenue quand la position et les déformations du modèle atteignent un minimum de la fonctionnelle d'énergie. Comme les snakes, la fonction coût des modèles paramétriques est la somme d'une énergie interne, représentant les contraintes géométriques du modèle (distances, symétrie...) et d'une énergie externe, qui est définie à partir des données de l'image. Là où les contraintes internes n'agissent que localement avec les snakes, l'énergie interne des modèles paramétriques limite explicitement les déformations possibles du modèle en intégrant des informations globales sur la forme attendue de l'objet. La liberté des déformations est restreinte car l'énergie n'est pas minimisée par rapport aux points du modèle, mais par rapport aux paramètres des courbes. Ce type de paramétrisation permet notamment d'être plus robuste vis-à-vis d'occultations partielles.

L'utilisation des modèles paramétriques requiert la définition du modèle (choix des courbes paramétrées) en fonction de la forme de l'objet d'intérêt. Le choix du modèle est un compromis entre la liberté de déformation souhaitée et la complexité algorithmique.

Le principal avantage des modèles paramétriques sont les contraintes géométriques qui imposent un assortiment de formes admissibles pour le résultat de la segmentation. Ceci permet d'éviter des déformations trop libres et le contour obtenu après convergence est cohérent avec le modèle prédéfini. Cependant, si la limitation des déformations est trop contraignante, le modèle devient trop rigide. Ceci peut être particulièrement délicat pour des objets hautement déformables tels que la bouche par exemple. De la même manière que pour les snakes, l'initialisation est une étape cruciale à cause de l'étape de minimisation de l'énergie.
Les algorithmes basés sur les modèles paramétriques pour l'analyse labiale nécessitent la définition de trois étapes : le choix du modèle pour la description des contours, l'initialisation et l'optimisation du modèle prenant en compte des informations appropriées et calculées à partir de l'image.

Dans le chapitre 3, plusieurs travaux utilisant les modèles paramétriques pour détecter les contours des lèvres seront présentés ([Hennecke, 1994]; [Zhang, 1997]; [Yin, 2002]; [Wu, 2002]; [Zhiming, 2002]; [Werda, 2007]...).

2.5. Bilan de l'état de l'art sur les méthodes de segmentation des contours des lèvres

Dans ce chapitre, nous avons effectué un état de l'art des méthodes utilisées dans la littérature pour l'analyse labiale. Les approches ont été séparées en deux classes : les approches orientées région et les approches orientées contour.

Concernant les techniques basées région (cf. section 2.3), nous avons mis en évidence que généralement ces méthodes ne permettaient pas d'obtenir une grande précision. Les méthodes de seuillage ou de classification fournissent une segmentation des lèvres qui est bruitée et où les contours ne sont pas particulièrement lisses et exacts. Ce type d'approche peut être utilisé pour détecter la région de la bouche ou obtenir des paramètres labiaux. Les modèles statistiques (ASM et AAM) proposent une plus grande fidélité des contours, mais ne fournissent pas encore des résultats assez précis pour des applications de lecture labiale par exemple. Les contours étant représentés par un ensemble de points plus ou moins espacés, le résultat ne semble généralement pas naturel (cf. Fig. 2.19). Cependant, il est possible d'interpoler ces points de façon non linéaire pour améliorer le côté non naturel. De plus, la segmentation est très dépendante de la base d'apprentissage et chaque cas à traiter doit avoir été pris en compte au cours de l'entraînement; ce qui pour les lèvres n'est pas forcément évident du fait de leur forme hautement déformable.

Fig. 2.19. Exemples de résultats de segmentation en utilisant un ASM et un AAM [Gacon, 2005].

Dans la section 2.4 nous avons brièvement présenté les modèles déformables (les contours actifs et les modèles paramétriques). Ces approches permettent d'obtenir des résultats rapides (coût de calcul limité) et précis. La difficulté vient du bon choix du modèle à déformer, de son initialisation et de son optimisation.

Dans le cadre de cette thèse, nous avons besoin d'extraire les contours extérieur et intérieur des lèvres et de les suivre dans des séquences d'images pour une application de lecture labiale et de maquillage virtuel. L'algorithme doit donc être rapide et précis. Nous nous sommes donc orientés vers des méthodes de type contour. Le chapitre suivant présente un état de l'art détaillé des méthodes basées sur les contours actifs et les modèles paramétriques dans le contexte de la segmentation des contours des lèvres.

CHAPITRE 3

Les modèles déformables : Etat de l'art

Dans la perspective de développer un algorithme de détection des contours des lèvres qui soit à la fois rapide et précis, nous nous sommes intéressés aux modèles déformables. Les contours actifs et les modèles paramétriques ont été brièvement présentés dans le chapitre précédent, où nous avons insisté sur leur définition et leurs propriétés respectives.

Dans ce chapitre, nous effectuons un état de l'art complet des travaux utilisant les modèles déformables pour faire de l'analyse labiale.

Dans la section 3.1, nous énumérons plusieurs travaux basés sur les contours actifs. Ces différentes approches sont classées selon le type de snake utilisé, la méthode d'initialisation choisie et les énergies externe et interne proposées.

Le même type d'étude est proposé dans la section 3.2 pour les modèles paramétriques. Nous présentons notamment les modèles et les courbes paramétrées les plus employés pour la détection des contours labiaux.

Finalement, nous terminons ce chapitre par une conclusion sur ces deux types d'approche et nous introduisons la démarche que nous avons décidée de suivre pour nos travaux.

3.1. Les contours actifs

Les contours actifs ont été particulièrement étudiés dans le cadre de la segmentation des lèvres pour leur simplicité et leurs propriétés de déformation. L'objectif consiste toujours à faire évoluer une courbe initiale vers un contour en minimisant une fonction d'énergie régie par des contraintes internes, liées à la géométrie de la courbe (courbure et tension), et externes, liées aux données saillantes de l'image (contours, zones claires ou sombres). Dans cette section, nous présentons les différents snakes, les méthodes d'initialisation des courbes proposées et les termes d'énergie considérés dans plusieurs travaux traitant de l'extraction des contours de la bouche.

3.1.1. Les différentes formes de snake utilisées pour la segmentation labiale

Depuis l'introduction des contours actifs par Kass *et al.* [Kass, 1987], différents modèles de snake ont été proposés pour réduire la dépendance de la segmentation vis-à-vis de l'initialisation ou pour améliorer leur force de convergence.

Wu *et al.* [Wu, 2002] ont développé un snake GVF basé sur le flux du vecteur gradient (Gradient Vector Flow; [Xu, 1998]). Le GVF est calculé comme étant une diffusion des vecteurs gradient sur l'image. Le champ résultant permet d'initialiser le snake relativement loin de l'objet à segmenter. De plus, contrairement à la définition standard, le contour actif peut être attiré par des zones concaves. Cependant, ce type de snake demande un temps de calcul très long, ce qui est pénalisant pour des applications temps réel.
Seguier *et al.* [Seguier, 2003] utilisent des snakes dit génétiques. Ils associent des algorithmes génétiques et des contours actifs pour surmonter les problèmes liés à l'initialisation et empêcher que le snake soit attiré par un minimum local.
Shinchi *et al.* [Shinchi, 1998] proposent un snake SACM (**S**ampled **A**ctive **C**ontour **M**odel) pour extraire le contour extérieur des lèvres. Les contours actifs standards convergent en minimisant une énergie, alors que les snakes SACM sont contrôlés par des forces appliquées sur les points de la courbe. Cette modification permet une extraction des contours beaucoup plus rapide.
Les B-snakes sont des contours actifs utilisant des B-splines. Ce type de snake a été développé dans [Blake, 1995]. Les B-splines sont des courbes fermées définies par plusieurs points de contrôle. Elles permettent de représenter une large variété de courbes et elles possèdent des propriétés de continuité intéressantes. En effet, les points des B-snakes sont reliés par des fonctions polynomiales incluant des propriétés de régularité intrinsèques. En conséquence, il est possible de n'utiliser que le terme d'énergie externe pour la déformation de la courbe et le coût de calcul en est allégé. Les B-snakes sont également utilisés pour la segmentation des contours des lèvres dans [Kaucic, 1998] et [Wakasugi, 2004].

3.1.2. Initialisation et choix des courbes

Pour éviter que le contour actif ne se bloque sur un minimum local, la courbe initiale doit être positionnée très proche des contours de la bouche. De plus, comme les snakes ont tendance à se contracter sur eux-même, la courbe initiale doit être placée à l'extérieur et autour de la bouche. Les paramètres du snake doivent être suffisamment bien choisis pour permettre de passer par dessus les contours parasites tout en ne dépassant pas le contour recherché. Plusieurs approches ont été proposées pour l'initialisation des contours actifs dans le contexte de la segmentation des lèvres. Elles se décomposent essentiellement en trois étapes : recherche de la zone de la bouche, choix du type de courbe initiale et positionnement du modèle.

3.1.2.a. Détection de la région d'intérêt de la bouche

De manière générale, la localisation de la bouche est réalisée à l'aide d'une des deux méthodes

suivantes : méthodes utilisant la projection de l'intensité ou du gradient de l'intensité et méthodes basées sur l'analyse d'informations couleur.

Dans ([Radeva, 1995]; [Delmas, 1999]; [Delmas, 2002]; [Kuo, 2005]), les projections de l'image intensité ou du gradient de l'intensité servent à la détection de la zone des lèvres. Dans [Radeva, 1995], la position de la bouche est trouvée en analysant les projections horizontales et verticales du plan intensité couplées à des contraintes de symétrie. Dans ([Delmas, 1999]; [Delmas, 20002]; [Kuo, 2005]), la position verticale de la bouche est détectée en accumulant verticalement les pixels sombres dans chaque colonne de l'image (cf. Fig. 3.01.a). Ensuite, les commissures des lèvres sont extraites en chaînant les pixels les plus sombres et en détectant les premiers sauts de la chaîne ainsi obtenue. La projection verticale du gradient de l'intensité donne les limites haute et basse de la bouche (cf. Fig. 3.01.b).

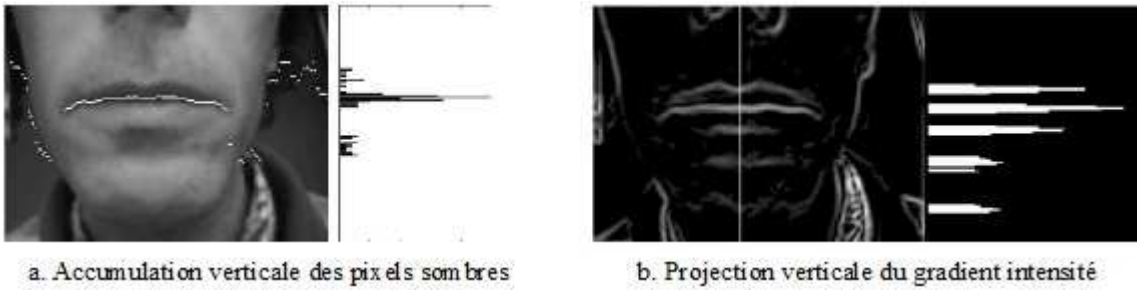


Fig. 3.01. Localisation de la bouche par projections verticales [Delmas, 1999].

Une autre approche consiste à déterminer la région d'intérêt à l'aide d'une analyse de la chrominance. Nous avons déjà présenté de telles méthodes bas niveau dans la section 2.3.1. Dans le cadre des contours actifs, nous pouvons mentionner également les travaux suivant qui utilisent l'espace couleur *YUV* ([Seguier, 2003]; cf. Fig. 3.02) ou l'espace couleur *HLS* ([Shinchi, 1998]; [Sugahara, 2000]; [Sasaki, 2004]).

Fig. 3.02. Détection de la région bouche [Seguier, 2003].

Des opérations de niveau supérieur peuvent être employées (cf. section 2.3.2). Par exemple, dans [Beaumesnil, 2006], une classification basée sur les k-means et utilisant le plan teinte *U* (cf. Eq. 2.02) découpe l'image en trois classes : lèvres, peau et fond (cf. Fig. 3.03).

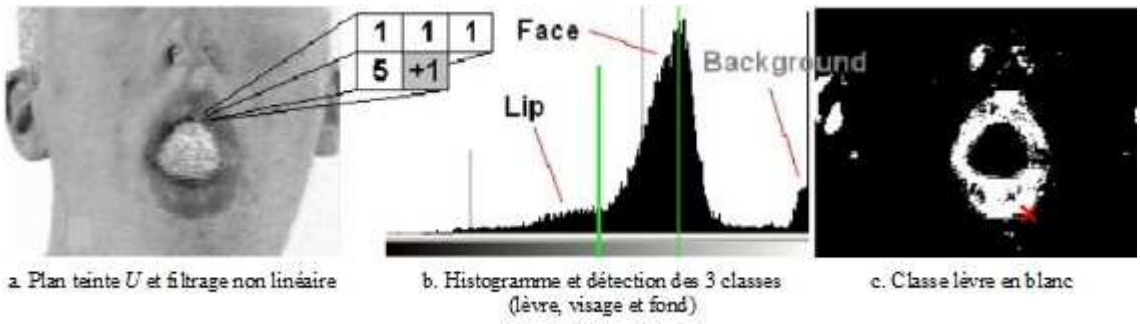


Fig. 3.03. Localisation de la bouche par classification [Beaumesnil, 2006].

3.1.2.b. Positionnement de la courbe initiale

Lorsque la bouche est localisée, la courbe initiale du snake doit être placée très près des contours des lèvres. La position est déterminée à partir d'une boîte encadrant la bouche, de la région elle-même ou de points clefs situés sur la bouche. Les méthodes proposées dans la littérature peuvent varier suivant que la segmentation est réalisée sur une seule image (méthode statique) ou dans une séquence vidéo (suivi des contours).

Dans le cas d'algorithmes statiques, le contour actif initial est placé directement sur une boîte encadrant la bouche ([Seo, 2003]; [Kuo, 2005]). Dans [Delmas, 1999]), les points du snake initial sont positionnés sur les contours extérieur et intérieur des lèvres. Les contours des lèvres sont choisis parmi les contours les plus forts de la région de la bouche à l'aide de la position des commissures.
Cependant, pour être proche des contours à segmenter, l'initialisation peut être réalisée à partir d'un modèle composé de plusieurs courbes; le modèle étant placé en prenant en compte des caractéristiques de la bouche ([Horbelt, 1995]; [Chiou, 1997]; [Delmas, 2002]). Dans [Horbelt, 1995], un modèle composé de plusieurs ellipses est initialisé à partir de la région de la bouche (cf. Fig. 3.04). Chiou *et al*. [Chiou, 1997] positionnent un petit cercle au centre de la bouche. Les points du cercle sont régulièrement espacés et leur position évolue en faisant varier la longueur des vecteurs radiaux. Dans [Delmas, 2002], l'extraction préalable des points extrêmes verticaux de la bouche et des commissures permet de positionner un modèle composé de quartiques, qui est échantillonné pour créer les points du snake.

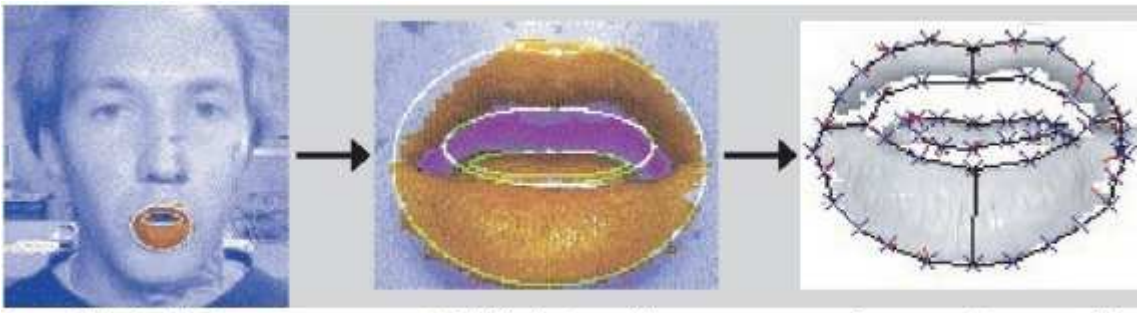


Fig. 3.04. Initialisation du snake [Horbelt, 1995].

Les travaux, qui ont pour but la détection des contours extérieur et intérieur des lèvres, peuvent utiliser la position du snake extérieur (construit pour segmenter le contour extérieur) pour initialiser le snake intérieur ([Beaumesnil, 2006]; [Seyedarabi, 2006]). Dans [Beaumesnil, 2006], le snake extérieur est initialisé par deux courbes cubiques calculées à partir d'une boîte encadrant la bouche et de la position des coins de la bouche. Le contour actif intérieur initial est déterminé en contractant le résultat de la convergence du snake extérieur par un changement d'échelle non isotrope par rapport au centre de la bouche et aux épaisseurs des lèvres. Dans le même genre d'idée, une forme ovale autour des lèvres est utilisée pour le snake extérieur, et le contour actif intérieur initial est donné par la convergence du premier snake [Seyedarabi, 2006].

Lorsque l'objectif est de suivre les contours des lèvres dans une séquence vidéo, l'initialisation du snake dans l'image courante est réalisée en utilisant des informations sur la forme des lèvres obtenues dans l'image précédente. Par exemple la position du contour actif final de l'image précédente est utilisée directement comme initialisation dans [Seyedarabi, 2006], alors qu'une dilatation de 20% est effectuée dans [Kuo, 2005].
Une autre approche consiste à initialiser le snake en suivant la position de quelques points clefs ([Delmas, 2002]; [Beaumesnil, 2006]) ou par *template matching* ([Barnard, 2002]; [Wu, 2002]; [Seo, 2003]).

Dans la première image de la séquence, il est possible de faire des hypothèses pour simplifier l'étape d'initialisation : soit les contours sont supposés être connus ([Pardas, 2001]), soit la bouche est considérée comme étant fermée ([Beaumesnil, 2006]; [Seyedarabi, 2006]).

3.1.3. Contraintes d'énergie du contour actif

Une fois que la position initiale du contour actif est trouvée, la minimisation de l'énergie fonctionnelle permet de déformer les points du snake pour qu'il se positionne sur le contour recherché. La convergence est réalisée lorsque l'énergie du snake est minimale. La spécification de la fonction d'énergie est une étape essentielle et de multiples définitions ont été proposées pour les applications d'analyse labiale. En plus des termes interne E_{int} et externe E_{ext}, des forces additionnelles liées à l'application elle-même ou le choix du modèle pour le snake peuvent compléter l'énergie fonctionnelle.

3.1.3.a. Contraintes de régularisation et énergie interne

L'énergie interne impose des contraintes locales et elle contrôle l'allure de la courbe à l'aide de deux termes : α ajuste la tension et β contrôle la courbure. α et β sont des constantes et elles sont fixées de manière heuristique.

Les contours actifs sont des algorithmes à forme libre, donc ils utilisent peu d'information sur la forme possible de l'objet. En ce qui concerne les lèvres, il est tout de même possible de prendre en considération des spécificités liées à la forme de la bouche. Par exemple, une difficulté mise en avant dans de multiples travaux concerne la convergence des snakes au niveau des coins de la bouche. Les commissures se trouvent dans des régions floues où le gradient est faible, et elles sont situées sur des contours concaves. Pour surmonter ce problème, certaines études suggèrent de modifier l'espacement des points du snake ou bien les valeurs des coefficients de tension et de courbure au niveau des coins de la bouche.

Dans [Seyedarabi, 2006], les points du snake initial sont espacés régulièrement, à l'exception de ceux situés dans la région des coins où la densité de points est plus élevée. Dans [Delmas, 1999], les points du snake sont réorganisés pendant les phases de déformation pour toujours être espacés régulièrement le long de la courbe. Dans le même genre d'idée, Pardas *et al.* [Pardas, 2001] introduisent un terme dans la définition de l'énergie interne qui force les nœuds du snake à conserver le même espacement entre eux.
Dans ([Delmas, 1999]; [Pardas, 2001]; [Seguier, 2003]), le coefficient de rigidité β n'est pas constant le long de la courbe. Sa valeur est plus importante au milieu de la bouche et elle est nulle aux points correspondant aux commissures. Ce choix est motivé par le fait que la courbure est minimale au milieu et maximale au niveau des coins. La même propriété est exploitée dans [Kuo, 2005], où α et β sont réduits près des commissures.
Radeva *et al.* [Radeva, 1995] imposent des contraintes symétriques par rapport à la ligne verticale passant au milieu de la bouche pendant les déformations de la courbe.
Pour réduire l'influence des coefficients α et β, certains travaux se servent de modèles de forme pour régulariser la courbe. Wu *et al.* [Wu, 2002] proposent d'utiliser un modèle composé de deux paraboles situées sur les contours préalablement trouvés par un détecteur de Canny. Les deux paraboles sont floutées et une contrainte additionnelle est obtenue avec un opérateur de gradient. De la même manière, nous avons vu que l'utilisation des B-splines permettait de ne pas avoir à définir d'énergie interne ([Kaucic, 1998]; [Wakasugi, 2004]).

3.1.3.b. Contraintes liées à l'image et énergie externe

L'énergie externe, liées aux données de l'image, conduit le snake vers les caractéristiques intéressantes de l'image telles que les lignes ou les contours. La définition la plus simple de l'énergie externe pour les lignes est l'image intensité, qui, selon le signe de l'énergie, force le contour actif à se diriger vers les lignes sombres ou claires. Le gradient de l'intensité est généralement utilisé pour attirer le snake vers les contours des lèvres.

Dans [Seyedarabi, 2006], l'énergie externe est définie comme étant la somme de l'image intensité et du gradient intensité. Cependant, certaines parties du contour de la bouche ne possèdent pas particulièrement un gradient intensité fort. De plus, des contours parasites, notamment à l'intérieur de la bouche, peuvent être accentués par le gradient intensité et perturber le bon déroulement de la convergence. Par exemple dans [Radeva, 1995], il est mis en évidence que la convergence d'un contour actif au niveau du contour extérieur bas des lèvres est difficile et dépend fortement des conditions d'illumination. La lèvre inférieure peut être vue comme une zone sombre, une zone claire ou une combinaison de ces deux possibilités, selon la position de la source de lumière. Pour résoudre ce problème, trois snakes sont définis avec trois gradients intensité différents correspondant aux trois cas. Le snake associé à l'énergie la plus faible après la convergence est choisi comme étant le meilleur candidat. Dans [Pardas, 2001], les auteurs utilisent également le gradient intensité, mais un second terme est ajouté à l'énergie externe. Ce second terme prend en compte l'erreur de compensation obtenue par une estimation de mouvement basée sur la technique de *block matching*. En conséquence, l'énergie externe est faible lorsque la texture est similaire à celle trouvée à l'image précédente.

Au lieu d'utiliser le gradient intensité, une carte des contours peut être calculée à partir d'un espace couleur. Beaumesnil *et al.* [Beaumesnil, 2006] proposent de calculer le gradient à partir d'une combinaison de la teinte et de la luminance. Cette approche est plus appropriée pour décrire les contours des lèvres plutôt que d'utiliser seulement la luminance.

Dans [Wu, 2002], la force extérieure utilise le flux de gradient du vecteur (GVF), ce qui a l'avantage d'accroître le champ de convergence du snake (cf. Fig. 3.05).

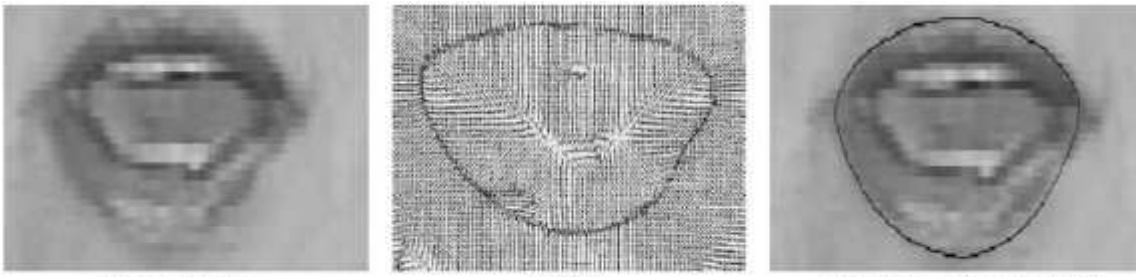


Fig. 3.05. Energie externe du snake GVF [Wu, 2002].

L'énergie externe peut être également formulée à l'aide d'une information couleur. Dans [Seo, 2003], chaque nœud du snake possède un patch couleur intérieur et un patch couleur extérieur. Les patchs sont décrits par des distributions gaussiennes utilisant les composantes rouge et verte. Dans le cas des pixels de la bouche, la composante intérieure est la couleur des lèvres et la composante extérieure est la couleur de la peau. Les nœuds évoluent avec une force extérieure qui prend en compte la différence entre les patchs et l'image originale. Kaucic *et al.* [Kaucic, 1998] accentue le contraste entre les lèvres et la peau par la projection de l'intensité des couleurs sur des axes déterminés par une analyse discriminante de Fisher. Cette méthode est plus efficace pour l'accentuation des lèvres qu'une transformation effectuée sur la composante teinte.

De multiples énergies externes de niveau supérieur ont été développées dans le cas de la détection des contours de la bouche. Par exemple, plusieurs études proposent différentes méthodes opérant sur des images binaires. Chiou *et al.* [Chiou, 1997] utilisent une image binaire obtenue par seuillage sur le rapport *R/G*. L'énergie externe est calculée sur chaque pixel de l'image binaire en utilisant un bloc de pixels binaires voisins. La valeur de l'énergie du pixel dépend de combien de pixels binaires sont proches du pixel courant. Seguier *et al.* [Seguier, 2003] définissent deux forces extérieures à partir d'images binaires de l'intérieur de la bouche et des lèvres. Le premier terme considère le nombre de pixels *N*

appartenant à l'image binaire de l'intérieur de la bouche et il est calculé comme étant le rapport *1/N*. Le second terme prend en compte la somme des valeurs de luminance des pixels appartenant aux lèvres et leur complément (la somme des pixels non lèvre).

Dans [Shinchi, 1998], les contraintes externes sont composées de trois forces. Une force de pression et une force d'attraction guident les points du snake vers le contour. Quand un point rencontre le contour des lèvres, une force de répulsion agit dans la direction opposée aux deux forces précédentes pour contre-balancer leur effet (cf. Fig. 3.06). Finalement, pour aider le snake à sortir des zones bruitées, un facteur de vibration est ajouté. Ce facteur agit perpendiculairement à la force attirant le snake vers le contour. La direction de la force de vibration s'inverse à chaque itération et ce mouvement en zigzag améliore l'aptitude du snake à surpasser les zones parasites. Dans [Barnard, 2002], une combinaison d'un contour actif et d'une méthode de *template matching* est utilisée pour extraire le contour extérieur de la bouche.

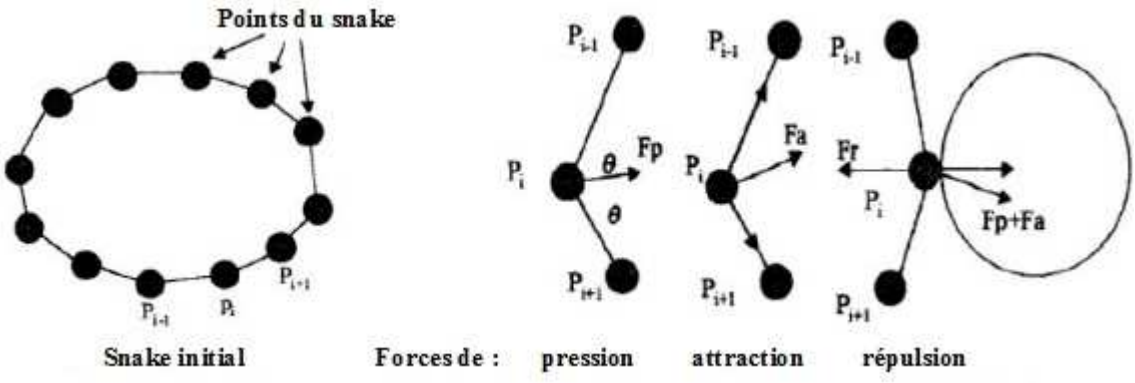


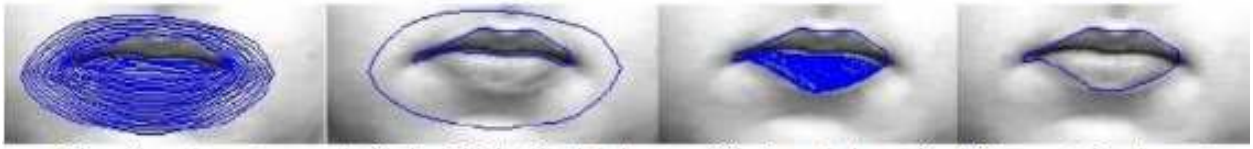
Fig. 3.06. Les 3 forces extérieures utilisées dans [Shinchi, 1998].

3.1.3.c. Forces extérieures additionnelles

L'énergie externe présentée dans le paragraphe précédent est définie pour guider le snake vers les caractéristiques saillantes de l'image, mais plusieurs minima locaux peuvent être également mis en valeur. En conséquence, des contraintes additionnelles externes prenant en compte l'application peuvent être utilisées pour rediriger le snake vers le minimum local désiré.

Par exemple, dans le cas de l'extraction des contours des lèvres, les coins de la bouche se situent dans une région où le gradient est faible et un contour actif est difficilement attiré dans leur direction. Plusieurs travaux proposent de détecter au préalable les commissures et d'ajouter des forces extérieures à la fonctionnelle d'énergie du snake pour soit attirer le contour actif vers les coins de la bouche pendant la phase de déformation, soit fixer les points du snake au niveau des commissures ([Liévin, 2004]; [Kuo, 2005]).

En prenant en compte la forme particulière de la bouche, les contours extérieur et intérieur des lèvres sont initialisés avec des courbes fermées lorsque la bouche est ouverte. En conséquence, la force ballon introduit par Cohen [Cohen, 1991] est fréquemment ajoutée dans l'expression de l'énergie du contour actif. La force ballon permet de gonfler ou dégonfler une courbe fermée comme un ballon. Cette opération permet de dépasser les contours des zones bruitées. La définition classique est utilisée pour gonfler un contour actif initial situé à l'intérieur de la bouche dans [Chiou, 1997] ou pour compenser le fait que les snakes ont tendance à se contracter sur eux-même dans [Delmas, 1999]. Seyedarabi *et al.* [Seyedarabi, 2006] emploient l'énergie ballon pour dégonfler un snake initial placé autour de bouche, afin d'obtenir le contour extérieur supérieur (cf. Fig. 3.07.a et b). Ensuite, seule la partie basse du snake évolue en utilisant une force ballon qui gonfle le contour, afin de détecter le contour extérieur bas.



a. Dégonflage du contour initial. Détection du contour extérieur supérieur des lèvres

b. Snake initial et final de la première phase

c. Gonflage de la partie basse du snake

d. Segmentation du contour extérieur des lèvres

Fig. 3.07. Méthode de convergence du snake proposée dans [Seyedarabi, 2006].

Dans [Kuo, 2005], une force ballon modifiée est proposée utilisant une équation de pression dépendante de la similarité des couleurs. Les paramètres d'un modèle gaussien de couleur sont ajustés en analysant le profil de la teinte le long de la verticale passant par le milieu de la bouche. Dans l'expression de la force ballon, un coefficient contrôle la puissance de gonflage ou de dégonflage en fonction de cette similarité. En d'autres termes, les régions ayant une teinte similaire donnent une pression positive, tandis que les autres régions fournissent une pression négative. Finalement, la force additionnelle permet de dégonfler un contour initial localisé autour d'une boîte encadrant la bouche. Dans [Beaumesnil, 2006], les auteurs forcent le contour actif à se dégonfler et à converger vers le centre de gravité de la région de la bouche.

3.1.4. Discussion sur les contours actifs

La principale difficulté lorsque l'on utilise les contours actifs est la phase d'initialisation. En ce qui concerne l'analyse labiale, l'initialisation est d'autant plus problématique que la configuration de la bouche peut créer plusieurs contours parasites accentués par le calcul de gradient. L'intérieur de la bouche contient différentes zones qui peuvent être très différentes (dents, cavité orale) ou très proches (gencives, langue) de l'aspect des lèvres. De plus, la présence de barbes, de moustaches ou de rides proches de la bouche peut également être une source de contours parasites. De manière générale, le contour extérieur bas des lèvres est difficile à extraire car l'aspect de ce contour varie beaucoup en fonction des conditions d'illumination.

Un autre inconvénient lié à la spécificité de la forme de la bouche est la convergence des snakes au niveau des commissures. Les coins de la bouche sont des régions concaves associées à un gradient souvent faible. En conséquence, les forces externes sont en général trop faibles pour compenser l'élasticité du contour actif. Le contour final peut de ce fait être trop rond autour des commissures et ne pas coïncider exactement avec leur position.

Cependant, les contours actifs sont une solution intéressante pour l'extraction des contours des lèvres, car ils offrent de grandes possibilités de déformation. Ceci est très utile dans la mesure où la forme des lèvres est hautement déformable. De plus, les snakes sont particulièrement appropriés aux calculs temps réel car leur implémentation est simple et rapide.

3.2. Les modèles paramétriques

Les modèles paramétriques sont des modèles déformables, qui, à l'instar des contours actifs, évoluent en minimisant une fonction d'énergie. La principale différence avec les contours actifs est qu'avec les modèles paramétriques, il est possible de prendre en compte *a priori* des contraintes géométriques de forme. Dans le cadre de l'analyse labiale, ils nécessitent d'effectuer trois phases déterminantes pour la qualité de la segmentation : 1) le choix du modèle pour la description des contours labiaux, 2) l'initialisation de la position du modèle dans l'image et 3) l'optimisation du modèle en fonction des caractéristiques de l'image.

3.2.1. Modélisation des contours de la bouche

La bouche est une caractéristique faciale hautement déformable et la première étape consiste à choisir un modèle paramétrique suffisamment flexible pour représenter fidèlement les contours des lèvres quelle que soit la forme visible dans l'image. Les lèvres peuvent prendre de multiples configurations qui sont autant de formes différentes (une bouche fermée ou largement ouverte, une grimace... cf. Fig. 3.08). Le modèle choisi doit donc pouvoir subir d'importantes déformations.

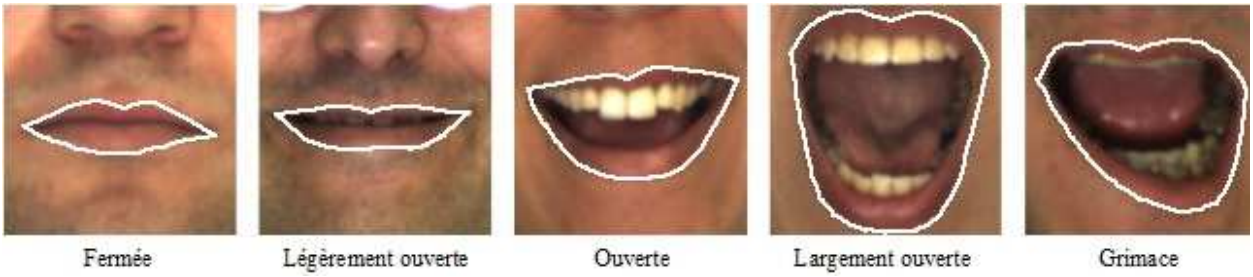


Fig. 3.08. Configurations possibles de la bouche [Martinez, 1998].

Plusieurs modèles paramétriques de lèvres ont été proposés depuis le premier modèle de Yuille *et al.* [Yuille, 1992]. Ils sont généralement composés de paraboles (polynôme de second degré défini par trois paramètres), de courbes cubiques (polynôme de troisième degré défini par quatre paramètres) ou des quartiques (polynôme de quatrième degré défini par cinq paramètres). Le choix des courbes utilisées dépend de la précision désirée et de la complexité admissible. Le type de courbe n'est pas nécessairement le même pour chaque partie des lèvres. Par exemple, le modèle représentant le contour extérieur supérieur de la bouche peut être plus complexe à cause de la présence de l'arc de Cupidon (forme en « V » visible au milieu du contour haut). En outre, des modèles différents peuvent être choisis suivant l'état de la bouche (ouverte ou fermée; [Zhang, 1997]; [Yin, 2002]) ou suivant la configuration de la bouche (par exemple, un modèle pour chaque cas : « bouche ouverte », « bouche relativement fermée » et « bouche bien fermée » [Tian, 2000a]). Dans cette section, nous commençons par présenter quelques modèles proposés pour le contour intérieur des lèvres, puis différents modèles pour le contour extérieur.

3.2.1.a. Modèles paramétriques des contours intérieurs

Deux aspects doivent être considérés : est-ce qu'un même modèle peut décrire à la fois une bouche ouverte et une bouche fermée? Est-ce que le modèle choisi doit rejoindre les coins de la bouche (et de ce fait, être relié au modèle du contour extérieur, cf. Fig. 3.09.b) ou doit-il être limité par des commissures « internes » (cf. Fig. 3.09.c)?

Fig. 3.09. Choix pour le modèle paramétrique intérieur.

Les courbes les plus utilisées pour la conception des modèles du contour intérieur des lèvres sont sans aucun doute les paraboles. Les coordonnées *(x, y)* d'une parabole peuvent être calculées par l'équation 3.01, où *h* est la hauteur de la courbe et *w* la largeur :

$$y = h\left(1 - \frac{x^2}{w^2}\right) \tag{3.01}$$

Dans [Yuille, 1992], le modèle relie les deux coins de la bouche avec deux paraboles qui sont superposées si la bouche est fermée (cf. Fig. 3.10). Ce modèle impose une contrainte de symétrie verticale qui n'est pas toujours vérifiée (cf. Fig. 3.10.c). Il y a 6 paramètres à régler, qui sont les coordonnées du centre *(xc, yc)*, l'angle d'inclinaison θ, la largeur de la bouche *w3+w4* (ici, *w3=w4*), la hauteur intérieure haute *h3* et la hauteur intérieure basse *h4*. Le même modèle intérieur est également utilisé dans ([Hennecke, 1994]; [Coianiz, 1996]; [Zhiming, 2002]), à la différence que les contours extérieur et intérieur des lèvres ne sont pas reliés aux coins de la bouche (cf. Fig. 3.10.d et e).

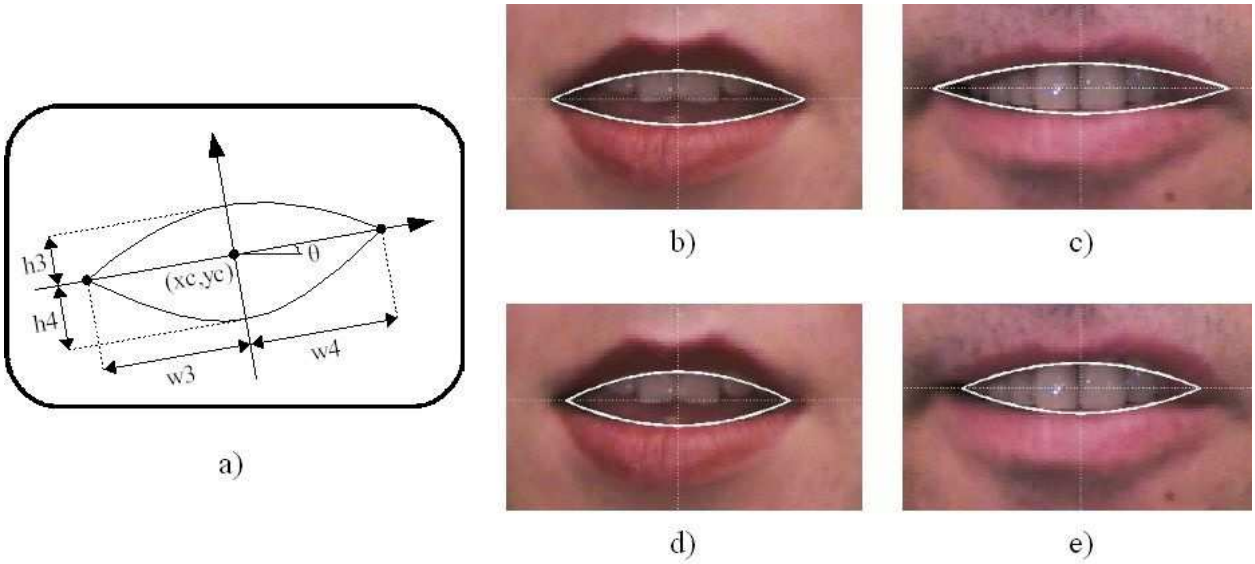


Fig. 3.10. a) Modèle intérieur à 2 paraboles. b) et c) utilisation des coins de la bouche. d) et e) utilisation des commissures internes.

Dans ([Zhang, 1997]; [Yin, 2002]), un modèle pour les bouches fermées composé d'une seule parabole (cf. Fig. 3.11) et un modèle pour les bouches ouvertes composé de deux paraboles (celui de [Yuille, 1992]) sont proposés. Dans [Wu, 2002], les auteurs développent un modèle intérieur associant aussi deux paraboles, mais contrôlé par seulement 4 paramètres, qui sont les distances gauche *d1*, droite *d2*, haute *j1* et basse *j2* définies entre les contours extérieurs et intérieurs des lèvres (cf. Fig. 3.24.a).

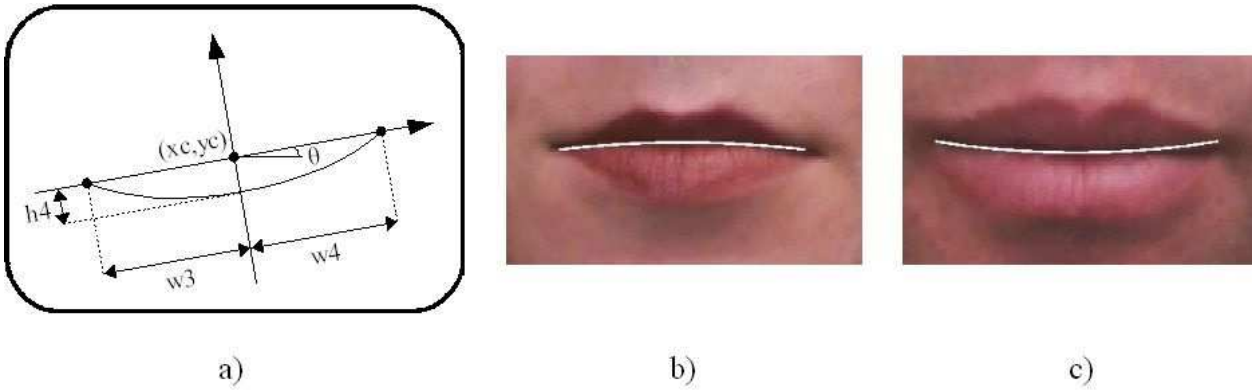


Fig. 3.11. a) Modèle intérieur à une parabole. b) et c) exemples de résultat.

Dans [Chen, 2006], le contour intérieur supérieur est représenté par deux paraboles (permettant une forme asymétrique) et le contour intérieur inférieur par une seule parabole. Pantic *et al.* [Pantic, 2001] font la même chose pour le contour haut, mais ils utilisent également deux paraboles pour le contour bas, ce qui fournit un modèle intérieur composé de quatre paraboles au total. Ce modèle est asymétrique car chaque côté de la bouche est traité séparément pour une meilleure représentation des contours labiaux, tout en gardant une certaine simplicité (6 paramètres, qui sont les 5 mêmes que ceux de [Yuille, 1992] plus la largeur *w4* qui prend une valeur différente de *w3*). Comme dans ([Hennecke, 1994]; [Coianiz, 1996]; [Zhiming, 2002]), *w3* et *w4* représentent les dimensions de l'intérieur de la bouche en utilisant des commissures internes.

Vogt [Vogt, 1996] et Malciu *et al.* [Malciu, 2000] proposent des modèles intérieurs composés de plusieurs nœuds (cf. Fig. 3.17). Dans [Vogt, 1996], le contour intérieur est seulement calculé lorsque la bouche est fermée et il est décrit par une courbe de Bezier composée de six points. Dans [Malciu, 2000], deux B-splines représentent les contours supérieur et inférieur des lèvres et chaque B-spline possède cinq nœuds.

Les modèles paramétriques des contours intérieurs des lèvres proposés dans la littérature sont majoritairement définis à partir de courbes d'ordre 2. Ceci sous-entend que le contour labial intérieur est relativement peu déformable et qu'il peut être modélisé par un modèle composé de paraboles. Les études plus récentes de [Pantic, 2001] et [Chen, 2006] montrent cependant que le modèle intérieur ne doit pas être trop simple et qu'il est notamment nécessaire de construire un modèle asymétrique pour obtenir une segmentation efficace.

3.2.1.b. Modèles paramétriques des contours extérieurs

En ce qui concerne l'extraction du contour extérieur des lèvres, le modèle du contour haut et celui du contour bas doivent être différents à cause de la présence de l'arc de Cupidon. Ainsi, de manière générale, les modèles extérieurs sont plus complexes que les modèles intérieurs présentés précédemment.

Le premier modèle extérieur, proposé par Yuille *et al.* [Yuille, 1992], utilise des quartiques. Les coordonnées *(x, y)* d'une quartique peuvent être calculées avec l'équation 3.02, où *h* est la hauteur de la courbe, *w* la largeur et le paramètre *q* contrôle la dérivée de la quartique.

$$y = h\left(1 - \frac{x^2}{w^2}\right) + 4q\left(\frac{x^4}{w^4} - \frac{x^2}{w^2}\right) \tag{3.02}$$

Le modèle de [Yuille, 1992] est composé de trois quartiques et il impose une contrainte de symétrie (cf. Fig. 3.12). Il y a 8 paramètres à régler, qui sont les coordonnées du centre *(xc, yc)*, l'angle d'inclinaison θ, la largeur de la bouche *w1+w2* (ici, *w1=w2* et elle est aussi égale à la largeur *w3* du modèle intérieur; cf. Fig. 3.10), la hauteur extérieure haute *h1* et la hauteur extérieure basse *h2*, le offset *a_off* du centre des quartiques et le paramètre *q* de l'équation 3.02. Le modèle complet (extérieur et intérieur) de Yuille *et al.* nécessite donc de régler 11 paramètres. Comme pour le modèle intérieur, la symétrie du modèle n'est pas toujours judicieuse pour décrire le contour extérieur des lèvres (cf. Fig. 3.12.c). Le même modèle extérieur est également utilisé dans ([Hennecke, 1994]; [Zhiming, 2002]) avec différentes valeurs entre la largeur externe de la bouche (2**w1*) et la largeur interne (2**w3*), ce qui donne un modèle complet à 12 paramètres. Dans [Yokogawa, 2007], le modèle est plus complexe car deux quartiques servent à représenter le contour extérieur bas.

De la même manière que pour la modélisation du contour intérieur, de multiples travaux suggèrent l'utilisation de paraboles ([Rao, 1995]; [Zhang, 1997]; [Tian, 2000a]; [Yin, 2002]; cf. Fig. 3.13). Il y a 6 paramètres à estimer, qui sont les coordonnées du centre *(xc, yc)*, l'angle d'inclinaison θ, la largeur de la bouche *w1+w2* (ici, *w1=w2*), la hauteur extérieure haute *h1* et la hauteur extérieure basse *h2*. La figure 3.13.c illustre le fait que de tels modèles sont généralement trop simplistes pour l'analyse labiale.

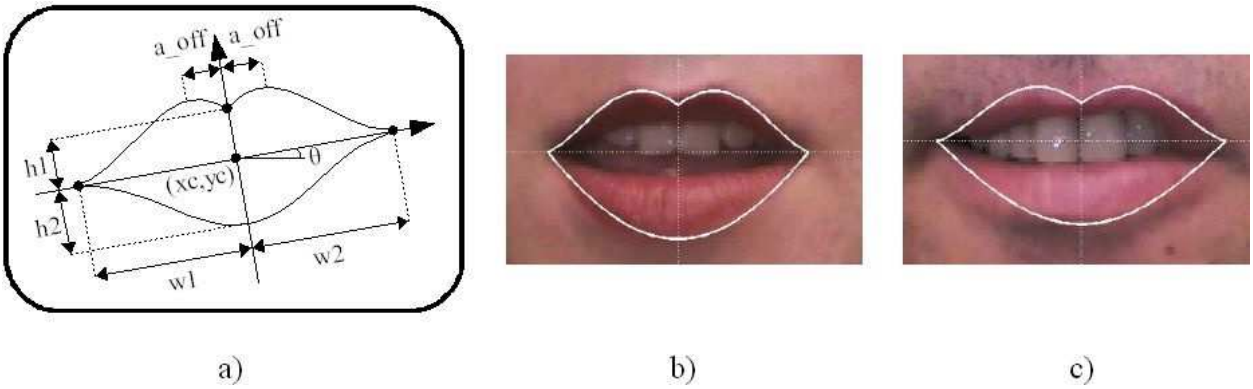


Fig. 3.12. a) Modèle extérieur à 3 quartiques. b) et c) exemples de résultat.

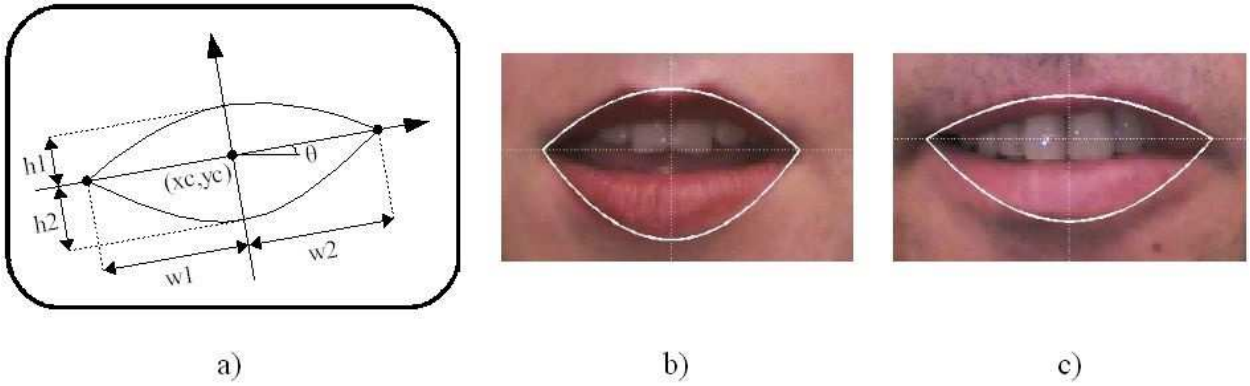


Fig. 3.13. a) Modèle extérieur à 2 paraboles. b) et c) exemples de résultat.

Dans ([Liew, 2000]; [Werda, 2007]), les deux formes paraboliques du modèle sont complétées par deux transformations de torsion et d'aplatissement, pour rendre la modélisation plus fidèle à la forme des lèvres. Dans [Pantic, 2001], quatre paraboles décrivent le contour extérieur de la bouche. Chaque côté de la bouche est traité séparément et le modèle est asymétrique (modèle extérieur à 7 paramètres : *xc*, *yc*, θ, *w1*, *w2*, *h1* et *h2*). Le modèle complet (extérieur et intérieur) possède donc 11 paramètres à régler. Toutefois, l'arc de Cupidon n'est pas correctement représenté, car les auteurs supposent que les points, où les dérivées des paraboles sont nulles, sont sur le même axe.

Dans [Coianiz, 1996], Coianiz *et al.* proposent également de modéliser le contour supérieur avec deux paraboles, mais leurs centres sont ajustés par le paramètre *a_off* (cf. Fig. 3.14). Cette approche amène à une meilleure description de l'arc de Cupidon. Yokogawa *et al.* [Yokogawa, 2007] utilisent le même modèle pour le contour haut, mais ils modélisent le contour bas avec deux paraboles au lieu d'une seule.

Les modèles extérieurs composés de paraboles sont intéressants pour déterminer certaines caractéristiques de la bouche, mais ils ne permettent pas d'obtenir des contours très précis soit à cause des contraintes de symétrie imposées, soit à cause de leur simplicité.

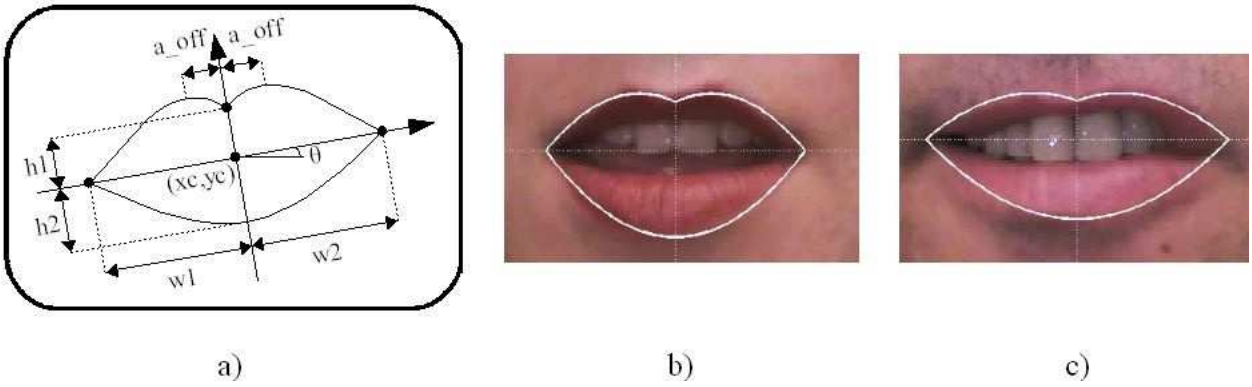


Fig. 3.14. a) Modèle extérieur à 3 paraboles. b) et c) exemples de résultat.

Dans [Eveno, 2003], les auteurs utilisent quatre courbes cubiques et une ligne brisée pour relier six points clefs et modéliser le contour extérieur des lèvres (cf. Fig. 3.15). Chaque courbe cubique nécessite la position de cinq points pour pouvoir être calculée avec la méthode d'approximation des moindres carrés (deux points pour les limites et trois points pour affiner la forme des cubiques). Le modèle est suffisamment flexible pour représenter des formes de lèvres très variables, et la ligne brisée est une description fidèle de la forme en « V » de l'arc de Cupidon. Bouvier *et al.* [Bouvier, 2007] utilisent le même modèle, à l'exception du contour extérieur bas qui est remplacé par deux courbes de Bézier.

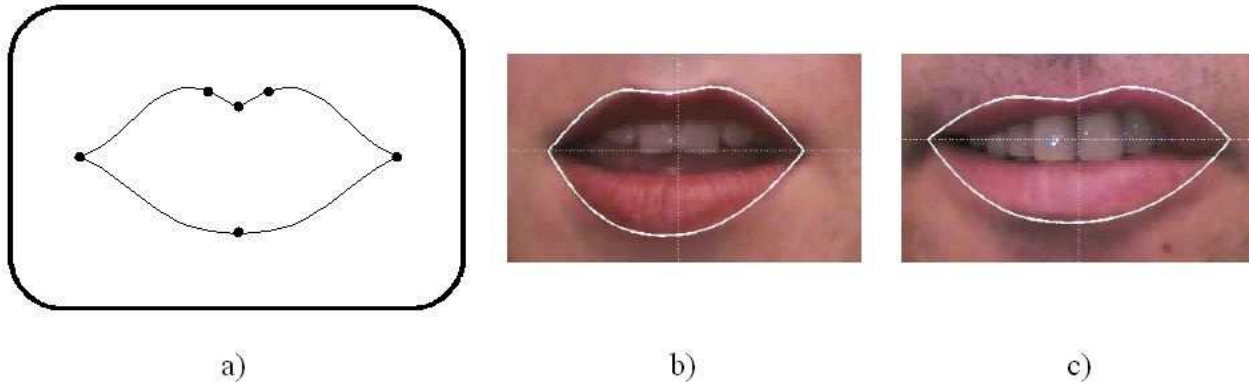


Fig. 3.15. a) Modèle extérieur à 4 cubiques [Eveno, 2004]. b) et c) exemples de résultat.

Salazar *et al.* [Salazar, 2007] montrent que le contour extérieur bas doit être modélisé par des courbes de quatrième degré (cf. Fig. 3.16.b) et le contour haut par un ensemble de trois courbes (deux fonctions de troisième ordre pour les côtés gauche et droit (cf. Fig. 3.16.c et d) et une fonction du quatrième ordre pour le milieu du contour (cf. Fig. 3.16.f)).

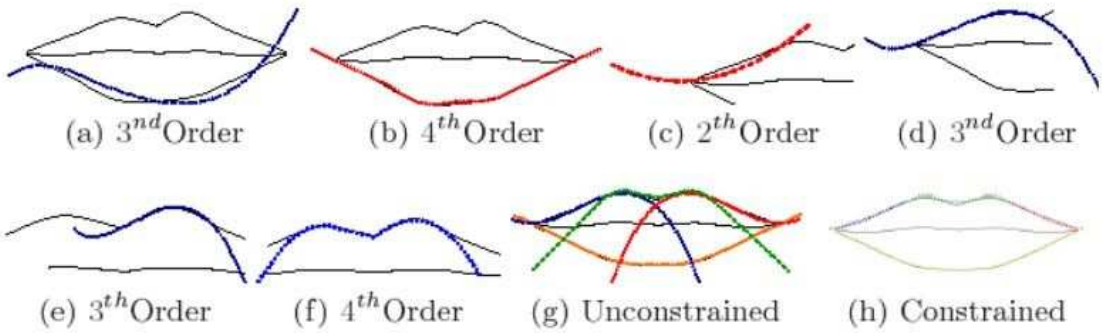


Fig. 3.16. Fonctions paramétriques pour la segmentation des lèvres [Salazar, 2007].

Le contour extérieur de la bouche est décrit par deux courbes de Bézier, composées de sept points pour le contour haut et six points pour le contour bas, dans [Vogt, 1996] et par des B-splines dans [Malciu, 2000] (cf. Fig. 3.17).

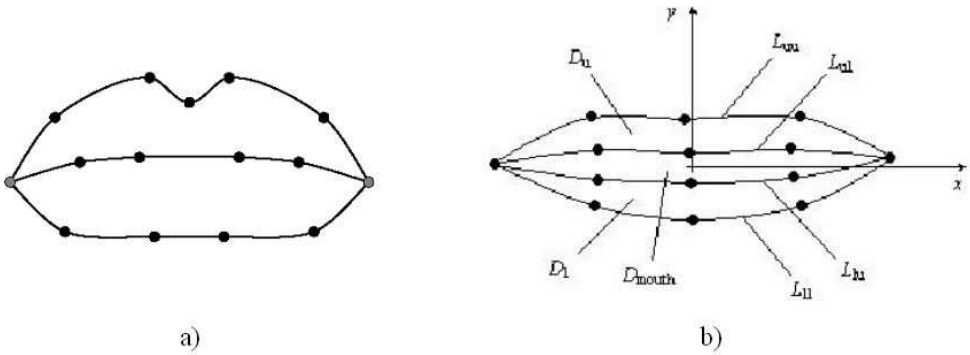


Fig. 3.17. a) Modèle à courbes de Bezier [Vogt, 1996], b) Modèle à B-splines [Malciu, 2000].

Cette section montre que, contrairement au contour intérieur, les modèles paramétriques composés de paraboles ne permettent pas de modéliser efficacement les contours extérieurs des lèvres. Par exemple, les courbes d'ordre 2 ne peuvent pas décrire précisément les déformations dues à la présence de l'arc de Cupidon sur le contour extérieur supérieur. Les modèles doivent être plus complexes que ceux proposés pour les contours intérieurs et les courbes utilisées sont généralement d'ordre 3 ou 4.

3.2.2. Initialisation des modèles paramétriques

Une fois que le modèle paramétrique des contours des lèvres a été choisi, il faut initialiser sa position dans l'image. Une détection de l'état de la bouche est nécessaire pour les approches qui utilisent des modèles différents selon que la bouche est ouverte ou fermée.

3.2.2.a. Sélection du modèle bouche ouverte ou bouche fermée

Certains travaux développent un modèle intérieur pour les bouches ouvertes et un modèle différent pour les bouches fermées ([Zhang, 1997]; [Yin, 2002]). Pendant la segmentation des contours, il est nécessaire de sélectionner le modèle adéquat et donc de détecter l'état de la bouche dans l'image traitée.

Dans la littérature, l'état ouvert ou fermé de la bouche est détecté en utilisant différentes approches, telles que l'analyse du gradient intensité [Zhang, 1997], les projections intégrales ([Pantic, 2001]; [Yin, 2002]), l'analyse colorimétrique [Chen, 2004] ou les méthodes de classification [Vogt, 1996].

Dans [Zhang, 1997], la région de la bouche est extraite à partir de la position des commissures des lèvres. Les contours présents à l'intérieur de la région d'intérêt sont détectés dans le plan *Y* de l'espace couleur *YCbCr* par un détecteur de contours et une série d'opérations morphologiques. Les intersections de ces contours avec la perpendiculaire de la ligne reliant les commissures des lèvres *W* fournissent le nombre de candidats possibles *P* pour les contours labiaux. La décision entre l'état ouvert ou fermé de la bouche est prise en considérant le nombre de candidats au dessus et en dessous de la ligne reliant les commissures. Si ce nombre est supérieur ou égal à deux pour chacun des deux cas, la bouche est supposée ouverte (cf. Fig. 3.18.c), sinon la bouche est fermée (cf. Fig. 3.18.a et b).

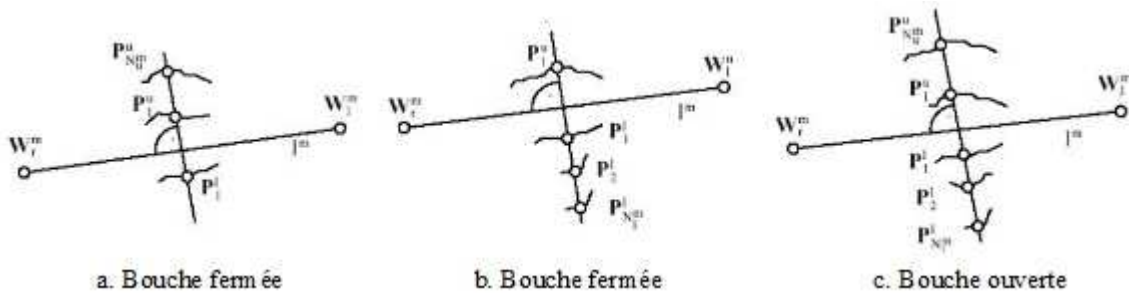


Fig. 3.18. Détection de l'état de la bouche dans [Zhang, 1997].

Pantic *et al.* [Pantic, 2001] transforment la composante teinte dans le domaine rouge pour obtenir la région de la bouche. La transformation est une opération de filtrage spécifique de la teinte [Gong, 1995]. La projection verticale de la composante du domaine rouge appliquée sur deux bandes verticales situées au centre de la région (cf. Fig. 3.19.a) donne le profil de la bouche. L'état de la bouche et l'épaisseur des lèvres sont déterminés à partir de ce profil (cf. Fig. 3.19.b et c). Dans [Yin, 2002], la projection intégrale de la composante teinte d'une bande verticale passant par le milieu de la bouche fournit un profil composé de deux pics. Si la largeur de la vallée est supérieure à deux pixels, la bouche est choisie ouverte, sinon elle est fermée.

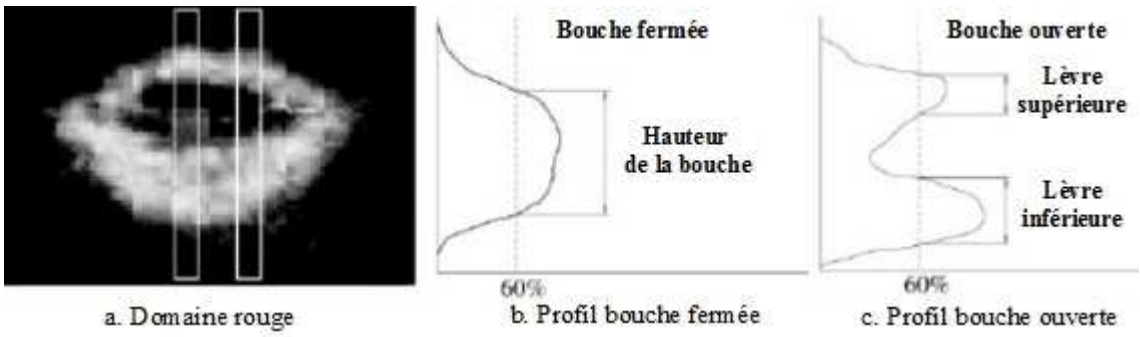


Fig. 3.19. Détection de l'état de la bouche dans [Pantic, 2001].

Chen *et al.* [Chen, 2004] appliquent plusieurs opérations de seuillage dans l'espace couleur *RGB* en prenant en compte la couleur de la peau et les pixels sombres. Les différents seuillages créent une image binaire des pixels sombres se trouvant entre la lèvre supérieure et la lèvre inférieure. Des opérations morphologiques du type érosion permettent de détecter les contours de la région sombre de l'intérieur de la bouche. Finalement, trois lignes traversant la bouche (une ligne verticale et deux lignes diagonales) déterminent la position de six points, qui sont les intersections des lignes et des contours de la zone sombre. En comparant la valeur des distances entre ces points clefs avec des valeurs de seuil, les auteurs arrivent à détecter l'état de la bouche.

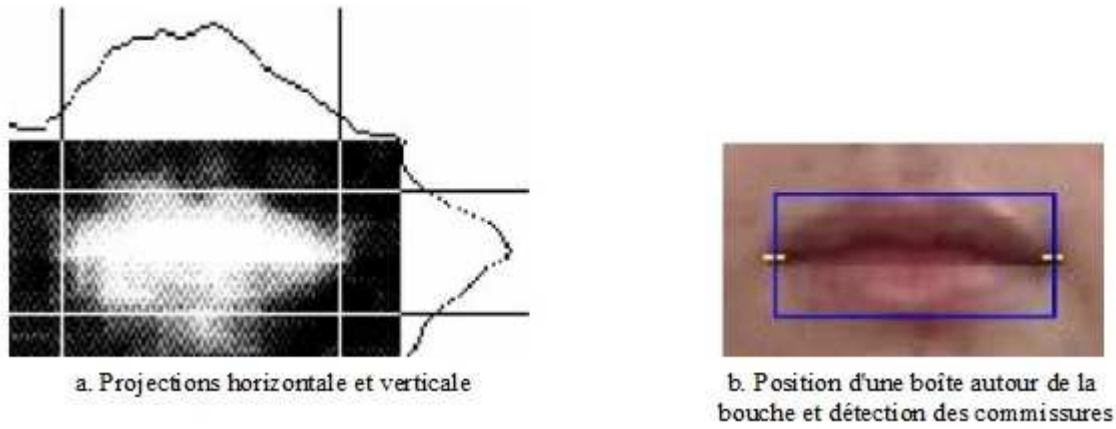
Dans [Vogt, 1996], un réseau de neurones est entraîné avec des valeurs références de teinte et d'intensité qui discriminent cinq classes : bouche fermée, bouche ouverte sans la présence des dents, bouche ouverte avec seulement les dents visibles à l'intérieur de la bouche, bouche ouverte avec les dents de la mâchoire supérieure et inférieure visibles et séparées par une région sombre (la cavité orale) et une bouche ouverte avec seulement les dents de la mâchoire supérieure visibles.

Les techniques employées pour détecter l'état de la bouche sont nombreuses et variées (projection, seuillage, réseau de neurones). La luminance et la teinte sont les deux composantes les plus utilisées car l'intérieur de la bouche est généralement sombre et la teinte permet de caractériser les lèvres (cf. section 2.1). La tâche est difficile et nous n'avons pas trouvé d'études consacrées uniquement à la détection de l'état de la bouche. Les méthodes présentées dans cette section ont donc été proposées dans le cadre de la segmentation des lèvres et pour chaque étude, aucune évaluation n'a été présentée.

3.2.2.b. Estimation des paramètres initiaux des modèles paramétriques

La convergence d'un modèle paramétrique est réalisée en deux étapes. Avant la phase d'optimisation et l'extraction des contours, le modèle doit être positionné approximativement dans l'image et près des contours labiaux. La position initiale peut être déterminée en détectant la région de la bouche ou en cherchant la position de plusieurs points clefs.

Les techniques les plus populaires pour localiser la région de la bouche sont les approches bas niveau de seuillage et de traitements morphologiques (cf. section 2.3.1). Certains travaux suggèrent d'utiliser des techniques de haut niveau. Dans [Rao, 1995], les régions lèvre et non-lèvre sont détectées avec un modèle HMM (**H**idden **M**arkov **M**odel) à l'intérieur d'une boîte encadrant la bouche, positionnée manuellement dans l'image. Zhiming *et al.* [Zhiming, 2002] positionnent une boîte autour de la bouche à partir de l'image précédente pour un suivi des contours des lèvres. Les contours labiaux sont accentués à l'aide de la transformée de Fisher dans la région de la bouche.
Dans [Coianiz, 1996], six points caractéristiques sont extraits de la région de la bouche, deux points où les contours extérieur et intérieur des lèvres se rejoignent (les commissures) et quatre points où la verticale du milieu de la bouche croisent les contours extérieur et intérieur. Les deux premiers points sont trouvés par une analyse de la chrominance et les quatre autres en utilisant l'information de luminance. Les amplitudes et les positions initiales du modèle paramétrique composé de paraboles sont initialisées à partir de ces six points. Dans [Pantic, 2001], des points clefs (les extrémités verticales) sont détectés par des projections intégrales verticales. Zhiming *et al.* [Zhiming, 2002] utilisent les projections horizontale et verticale de la région de la bouche pour extraire les commissures et positionner une boîte autour de la bouche très proche des contours (cf. Fig. 3.20). Dans [Werda, 2007], les projections de la composante de saturation permettent de détecter les coins de la bouche.



a. Projections horizontale et verticale

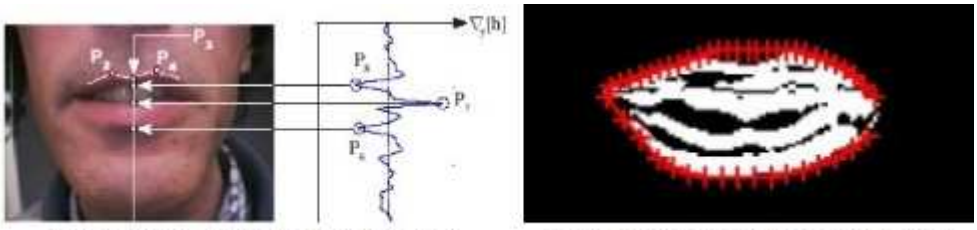
b. Position d'une boîte autour de la bouche et détection des commissures

Fig. 3.20. Initialisation du modèle paramétrique dans [Zhiming, 2002].

Comme nous l'avons vu dans la section précédente, Zhang *et al.* [Zhang, 1997] déterminent plusieurs candidats sur la ligne verticale passant par le milieu de la bouche. Les commissures ont été trouvées par une méthode de suivi de type *template matching*. Les candidats permettent d'initialiser des paraboles reliant les commissures.
On peut également citer les méthodes neuronales. Dans [Duffner, 2005], Duffner et al. localise plusieurs points sur le visage, dont quatre points sur le contour extérieur de la bouche, en utilisant une approche

basée sur des réseaux de neurones. Les réseaux de neurones permettent également de déterminer la position des commissures des lèvres dans [Muhammad Hanif, 2007].

L'utilisation des contours actifs peut être un moyen efficace d'initialiser les modèles paramétriques. Le résultat de la convergence du snake est lissé par le modèle paramétrique. Les contours actifs permettent d'augmenter la liberté de déformation des modèles paramétriques.
Dans [Jian, 2001], un contour actif est utilisé pour initialiser un modèle extérieur à deux paraboles. Un snake est également employé pour l'initialisation dans [Salazar, 2007]. Dans [Eveno, 2004], une série de points sur le contour extérieur supérieur et trois points clefs définissant l'arc de Cupidon sont détectés en faisant converger un snake (cf. Fig. 3.21.a). Bouvier *et al.* [Bouvier, 2007] calculent une carte des contours à partir de la région de la bouche. Un contour actif est appliqué sur l'image binaire pour obtenir plusieurs points sur le contour extérieur des lèvres (cf. Fig. 3.21.b).



a. Convergence d'un snake pour le contour supérieur haut (Eveno, 2004).

b. Convergence d'un snake pour le contour extérieur (Bouvier, 2007).

Fig. 3.21. Localisation de points en utilisant des snakes ([Eveno, 2004]; [Bouvier, 2007]).

En conclusion, nous pouvons dire que dans le cas où le but est de trouver un cadre autour de la bouche, la détection de la région d'intérêt peut être réalisée par de nombreuses méthodes de bas niveau (seuillage, opération morphologiques, projection) ou de haut niveau (modèle HMM). Si l'objectif est de déterminer la position de points clefs, ces mêmes méthodes sont peu précises et de nombreuses études proposent d'utiliser les contours actifs présentés dans la section 3.1.

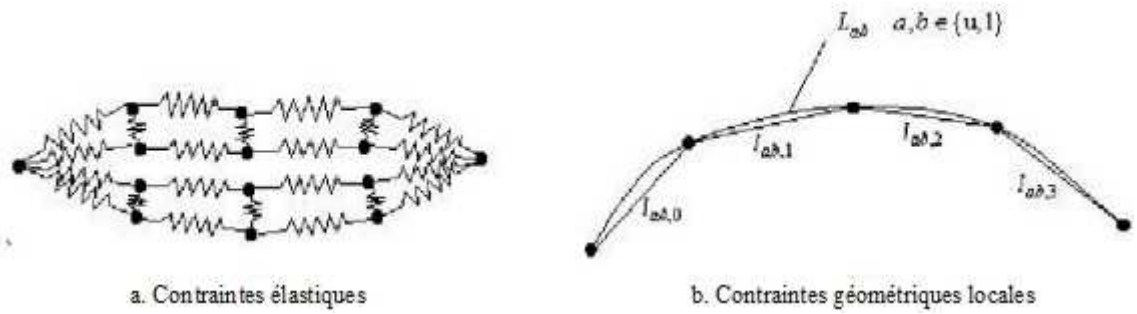
3.2.3. Optimisation des modèles paramétriques

Dans cette section, nous mentionnons différentes approches pour optimiser les modèles paramétriques et les faire converger sur les contours des lèvres. A l'instar des contours actifs, les modèles paramétriques évoluent de manière itérative en minimisant une fonction de coût composée d'un terme d'énergie interne et d'un terme d'énergie externe.

3.2.3.a. Energie interne pour la régularisation du modèle

L'énergie interne fixe des contraintes *a priori* sur la variabilité du modèle paramétrique. Dans [Yuille, 1992], l'énergie interne est un ensemble de contraintes telles que la symétrie du contour extérieur supérieur, le centrage de la bouche entre les commissures, une proportionnalité constante entre les épaisseurs des lèvres supérieure et inférieure et une cohésion empêchant un mouvement vertical trop large de la lèvre supérieure. Hennecke *et al.* [Hennecke, 1994] simplifient l'énergie interne de Yuille *et al.* en intégrant des notions de symétrie et de centrage directement dans l'élaboration du modèle paramétrique. Une contrainte sur la continuité temporelle des épaisseurs des lèvres est également introduite dans l'expression de l'énergie interne. De la même manière, Coianiz *et al.* [Coianiz, 1996] définissent des contraintes de pénalité qui représentent les formes admissibles. Par exemple, les largeur et hauteur du contour extérieur doivent avoir des valeurs plus élevées que les largeur et hauteur du contour intérieur. Dans [Mirhosseini, 1997], un potentiel est défini pour contrôler la forme de la bouche en mesurant les différences sur des distances caractéristiques des lèvres. Vogt [Vogt, 1996] propose une énergie interne qui stabilise la distance entre trois nœuds consécutifs du modèle à base de courbes de Bézier et contrôle les tensions entre deux points. La même idée est exploitée pour le modèle paramétrique composé de B-

splines dans [Malciu, 2000]. L'énergie interne est la combinaison de contraintes locales de symétrie et d'élasticité (cf. Fig. 3.22).



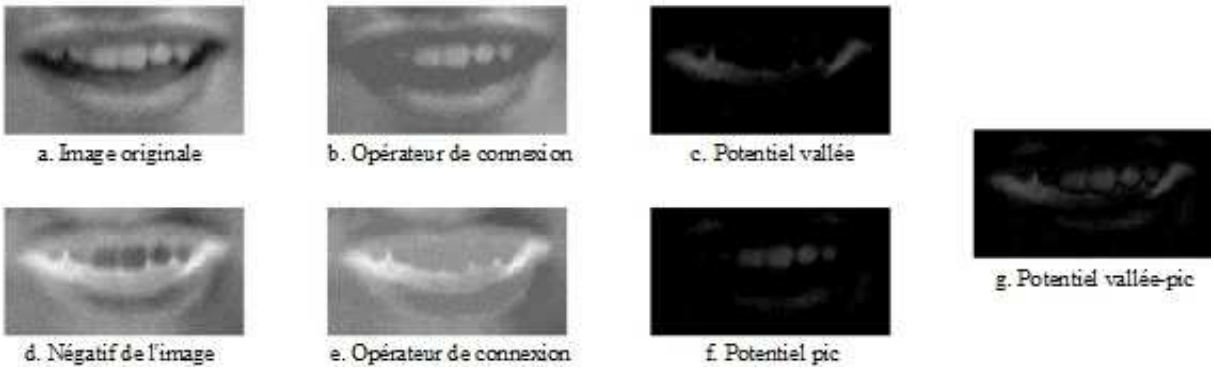
a. Contraintes élastiques b. Contraintes géométriques locales

Fig. 3.22. Energie interne proposée dans [Malciu, 2000].

3.2.3.b. Energie externe calculée à partir des données de l'image

De manière générale, les contraintes géométriques imposées par le modèle paramétrique suffisent et le modèle évolue seulement avec une énergie externe liée à l'image. L'énergie externe établit des contraintes d'interaction afin de maintenir une certaine consistance entre le modèle paramétré et les caractéristiques saillantes de l'image. L'information de l'image utilisée dans l'expression de l'énergie externe peut prendre de multiples formes.

La caractéristique la plus employée et la plus simple est l'image intensité et le gradient de l'intensité. Dans [Yuille, 1992], l'énergie externe est composée de trois champs énergétiques, le champ contour obtenu avec un opérateur gradient sur l'image intensité, le champ vallée des régions sombres de l'image et le champ pic des régions brillantes. Ces trois champs correspondent aux caractéristiques de l'image : les zones sombres, claires et de transition. Mirhosseini *et al.* [Mirhosseini, 1997] définissent un terme vallée et un terme issu du gradient intensité. Le même type d'information est exploité dans [Malciu, 2000]. Les informations pic et vallée sont combinées dans une même fonction potentielle en appliquant un opérateur de connexion dans l'image originale et le négatif de l'image (cf. Fig. 3.23). Le gradient intensité met en évidence les différents contours de la bouche et le potentiel vallée-pic accentue l'intérieur de la bouche composé de régions sombres (vallée) et claires (pic). Hennecke *et al.* [Hennecke, 1994] utilisent seulement la composante verticale du gradient intensité car les contours de la bouche sont essentiellement horizontaux. L'algorithme de Prewitt donne un gradient positif ou négatif selon que l'intensité est plus grande au-dessus ou au-dessous du contour. Le gradient approprié est appliqué selon les quatre contours des lèvres (les deux contours supérieurs extérieur et intérieur et les deux contours inférieurs extérieur et intérieur). Le gradient de l'image est également employé dans [Werda, 2007].



a. Image originale b. Opérateur de connexion c. Potentiel vallée g. Potentiel vallée-pic

d. Négatif de l'image e. Opérateur de connexion f. Potentiel pic

Fig. 3.23. Potentiel vallée-pic proposé dans [Malciu, 2000].

Le gradient intensité est principalement utilisé pour le contour extérieur des lèvres, mais il peut être faible pour certaines parties de l'image (contour extérieur bas par exemple) et il dépend des conditions d'illumination. Dans [Eveno, 2004] et [Bouvier, 2007], les auteurs suggèrent d'utiliser plutôt un gradient de chrominance. Le gradient calculé est une combinaison de la pseudo-teinte $\hat{H}$ et de la luminance pour le contour extérieur supérieur des lèvres et le gradient calculé directement sur la pseudo-teinte pour le contour inférieur.

Dans [Vogt, 1996], le gradient est obtenu à partir d'une carte de probabilité qui accentue les pixels lèvre. La carte est créée par un algorithme 2D *look up* appliqué sur la teinte et la saturation de l'espace couleur *HSI*. Ce gradient n'est utilisé que lorsque la bouche est ouverte, sinon il est remplacé par un gradient intensité pour le contour intérieur (le contour labial interne pouvant être vu comme une simple ligne sombre lorsque la bouche est fermée). Dans [Yokogawa, 2007], deux zones, dont la région de la bouche, sont détectées par un seuillage sur la teinte. Une opération *AND* permet de créer une image binaire décrivant les lèvres à partir des deux zones. Une carte des contours est construite avec un opérateur différentiel appliqué sur l'image binaire. L'énergie externe est une mesure de la différence entre le modèle paramétrique et les pixels contours de la carte.

Un modèle paramétrique de la bouche permet de séparer plusieurs régions telles que les lèvres, la peau et l'intérieur de la bouche. De multiples critères basés sur des informations couleurs ont été créés pour être minimaux lorsque les régions sont bien séparées et pour guider le déplacement des courbes du modèle ([Coianiz, 1996]; [Zhang, 1997]; [Pantic, 2001]; [Yin, 2002]; [Wu, 2002]).
Dans [Coianiz, 1996], trois zones sont définies : l'intérieur de la bouche, les lèvres et une zone d'épaisseur constante autour des lèvres. L'énergie externe prend en considération une mesure de la chrominance rouge dans ces trois régions et la maximisation (ou la minimisation selon le signe de l'énergie fonctionnelle) de l'énergie permet aux régions rouges de l'image d'être assimilées à la région lèvre. La même méthode avec la même information couleur est utilisée dans [Pantic, 2001]. Dans [Zhang, 1997], deux fonctions de coût sont proposées : une pour le modèle bouche fermée et une autre pour le modèle bouche ouverte. Elles sont définies avec le gradient intensité de l'image et pondérées par les moyennes et variances de la composante *Cr* de l'espace couleur *YCbCr* de chaque région du modèle paramétrique (la lèvre supérieure, la lèvre inférieure et, si la bouche est ouverte, l'intérieur de la bouche). La même approche est suivie dans [Yin, 2002], mais le critère est construit à partir de la teinte en considérant qu'une faible intra-variance et une forte inter-variance de la teinte existent entre ces trois régions (le critère est une somme des moyennes et écart-types de la teinte pour chacune des régions). Dans [Wu, 2002], le contour extérieur des lèvres est connu après la convergence d'un contour actif et le contour intérieur a été approximativement détecté par un modèle paramétrique. Pour améliorer la segmentation, deux régions doivent être distinguées : les lèvres et l'intérieur de la bouche. Les histogrammes des lèvres et de l'intérieur de la bouche sont appris avec une base d'apprentissage. Ensuite, une fonction coût estime les similarités entre les deux régions et l'apprentissage.

D'autres méthodes utilisent l'information couleur pour construire des cartes de probabilité aux lèvres (qui est une image représentant la probabilité de chaque pixel d'appartenir aux lèvres; [Rao, 1995]; [Liew, 2000]). Le critère a une valeur élevée quand la région incluse dans le modèle paramétrique possède les probabilités les plus fortes. Le critère a une valeur faible pour les régions extérieures aux lèvres. Dans [Liew, 2000], en assumant que la probabilité associée à chaque pixel est indépendante de la probabilité des autres pixels, la séparation optimale est obtenue quand la probabilité jointe des régions lèvre et non-lèvre est maximale.

3.2.3.c. Méthodes d'optimisation

Les algorithmes classiques pour la minimisation de l'énergie du modèle sont basés sur la méthode *downhill simplex* ([Yuille, 1992]; [Hennecke, 1994]; [Malciu, 2000]). La méthode *downhill simplex* est un algorithme d'optimisation non-linéaire qui minimise une fonction dans un espace à plusieurs dimensions. A partir d'une position initiale du modèle, la position évolue de manière itérative à travers les minima locaux en suivant les pentes les plus fortes. Les principaux inconvénients de cette méthode d'optimisation

sont le temps de convergence, qui augmente rapidement avec le nombre de variables de la fonction (dans notre cas, chaque nouvelle contrainte interne ou externe amène une nouvelle variable), et le fait que la solution n'est pas nécessairement le minimum global, le résultat dépend fortement de la proximité de la position initiale avec les contours recherchés. Yuille *et al.* [Yuille, 1992]) combinent la méthode *downhill simplex* avec un algorithme d'optimisation de type *coarse-to-fine*. La position générale du modèle (centrage par rapport à la bouche, calcul de l'orientation et de la largeur du modèle) est d'abord ajustée. Puis, la méthode *downhill simplex* permet d'obtenir une position plus précise des contours des lèvres. Ce déroulement en deux étapes est réalisé jusqu'à la convergence du modèle.

Dans [Vogt, 1996], les nœuds du modèle se déplacent aléatoirement dans une direction à chaque itération de la phase d'optimisation. Si la nouvelle position donne une meilleure solution (seulement la partie du modèle affectée par les nœuds courant est prise en compte), elle est conservée, sinon elle est rejetée.
Un algorithme de minimisation non-déterministe est utilisé dans [Coianiz, 1996] et le gradient conjugué, qui est une amélioration du *downhill simplex* et permet une optimisation plus rapide, est utilisé dans [Liew, 2000].

Une autre approche consiste à calculer les différentes positions possibles et à choisir la meilleure à l'aide d'un critère approprié. Dans [Zhang, 1997], les paraboles pour chaque contour candidat sont calculées et la fonction coût (cf. Section précédente) permet de déterminer la meilleure solution. Dans [Wu, 2002], les variations de quatre distances entre le contour extérieur connu et le contour intérieur des lèvres permettent de définir plusieurs positions possibles du modèle composé de paraboles. La meilleure position est choisie avec le critère introduit dans la section précédente (cf. Fig. 3.24.a).
Dans ([Eveno, 2004]; [Bouvier, 2007]), l'optimisation des courbes et la détection des commissures sont effectuées en une seule et même opération. Pour chaque côté de la bouche, un nombre fini de candidats possibles est testé pour la position de chaque coin de la bouche. Plusieurs courbes sont calculées en fonction des commissures testées et de points disponibles sur les contours des lèvres grâce à la convergence de contours actifs. La meilleure solution est choisie comme étant celle qui maximise le flux de gradient moyen à travers les courbes (cf. Fig. 3.24.b). Dans [Werda, 2007], toutes les positions possibles du modèle à l'intérieur de la boîte autour de la bouche sont calculées et la maximisation du flux de gradient permet de déterminer la meilleure configuration (cf. Fig. 3.24.c).

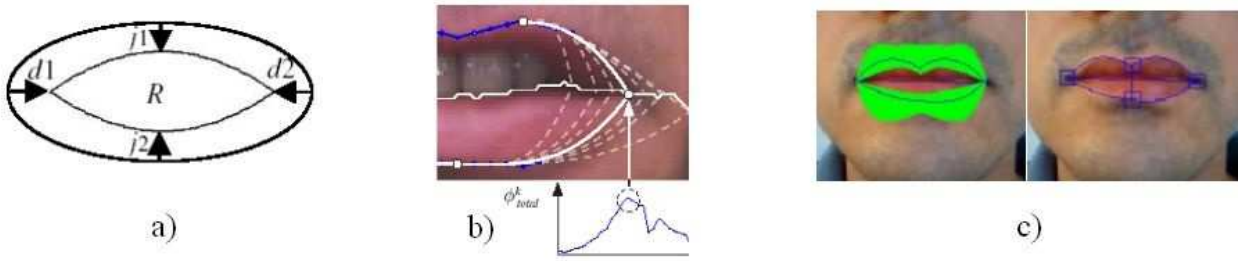


Fig. 3.24. Méthodes d'optimisation proposées dans la littérature. a) [Wu, 2002], b) [Eveno, 2004] et c) [Warda, 2007].

La position de plusieurs points est un moyen simple de calculer directement la position finale du modèle paramétrique. La convergence est très rapide, car elle est réalisée en une seule itération. Dans [Wark, 1998], les contours de la région de la bouche sont trouvés avec un détecteur de contour standard et ils sont échantillonnés pour fournir plusieurs points situés sur le contour extérieur des lèvres. La méthodes des moindres carrés permet d'optimiser le modèle paramétrique en fonction de la position de ces points.

Tian *et al.* [Tian, 2000a] suivent la position de quatre points clefs (les deux coins de la bouche et les deux extrémités verticales de la bouche) avec un algorithme de Lucas-Kanade [Tomasi, 1991]. Les paraboles du modèle sont complètement déterminées par la position des trois points sur le contour extérieur supérieur et des trois autres points sur le contour extérieur inférieur. Chen *et al.* [Chen, 2006] détectent les

contours de la région de la bouche avec un opérateur laplacien et les positions de plusieurs points sur le contour intérieur des lèvres sont déterminées. Un algorithme d'optimisation, appelé méthode d'interpolation aux plus proches voisins, est appliqué sur la carte des contours et il positionne le modèle paramétrique composé de trois paraboles sur le contour labial interne.

Une fois le modèle paramétrique choisi et initialisé, la dernière étape est la convergence du modèle. Dans cette section, nous avons présenté des méthodes d'optimisation guidées par des énergies internes et externes. L'énergie interne permet de régulariser le modèle en contrôlant l'épaisseur des lèvres ou en imposant des contraintes de symétrie par exemple. L'énergie externe est liée aux caractéristiques saillantes de l'image, et plus particulièrement aux contours. Ainsi, le gradient est généralement utilisé dans l'expression de l'énergie externe. Finalement, les méthodes d'optimisation (algorithme *downhill simplex* par exemple) utilisent les énergies internes et externes pour faire converger le modèle et extraire les contours des lèvres.

3.2.4. Discussion sur les modèles paramétriques

Comme pour les snakes, la principale difficulté des modèles paramétriques durant la phase de convergence est la possibilité de se retrouver sur un minimum local. En conséquence, l'initialisation est une étape cruciale et elle dépend de la précision de la détection de la région de la bouche ou de quelques points clefs.

Le choix du modèle doit être un bon compromis entre un modèle simple, qui amènerait une convergence rapide mais un résultat de segmentation peu précis, et un modèle complexe, qui fournirait une plus grande précision mais un temps de calcul long.
Dans le cadre de la segmentation des contours des lèvres, le modèle décrivant le contour supérieur doit être composé de courbes ayant un degré suffisamment grand, pour pouvoir représenter de façon réaliste l'arc de Cupidon. Il est également important d'avoir un modèle qui puisse décrire des formes de bouche asymétriques. De manière générale, le modèle intérieur est plus simple que le modèle extérieur, et l'hypothèse de relier les deux modèles par les commissures des lèvres est souvent adoptée. Cependant, ce choix n'est pas toujours judicieux, car lorsque la bouche a une forme ronde ou qu'elle est légèrement ouverte, les extrémités horizontales du contour intérieur se retrouvent proches du centre de la bouche.

Contrairement aux contours actifs, les modèles paramétriques introduisent une notion de connaissance *a priori* sur la forme globale de la bouche. Ainsi, le résultat de la segmentation après la convergence du modèle permet d'obtenir une forme admissible des contours. L'influence de perturbations locales est réduite par la minimisation de l'énergie globale. Nous avons vu que cette propriété pouvait être exploitée pour lisser les résultats obtenus avec un snake. La combinaison des contours actifs, utilisés pour fournir une estimation approximative des contours, et des modèles paramétriques, utilisés pour régulariser et lisser le résultat fournit les algorithmes les plus efficaces.

3.3. Conclusion et approche choisie

Dans la mesure où nous voulons obtenir un algorithme de segmentation et de suivi des contours extérieur et intérieur des lèvres qui soit à la fois précis et rapide, nous nous sommes orientés vers les modèles déformables et une approche basée contour.

Ce chapitre a mis en évidence les avantages et les inconvénients des contours actifs et des modèles paramétriques qui sont les deux types de modèles déformables les plus couramment employés pour une approche contour.
Les snakes permettent de reproduire des formes très réalistes car ce sont des algorithmes à forme libre, ce

qui est très intéressant pour la bouche qui présente des contours hautement déformables. Mais une trop grande flexibilité n'est pas aisée à contrôler et le résultat peut très vite être aberrant si l'initialisation est mal effectuée ou si des contours parasites sont présents autour des contours labiaux.
A l'inverse, les modèles paramétriques permettent d'obtenir un résultat toujours conforme aux attentes du concepteur du modèle, mais pour une bonne précision, il faut un modèle composé de plusieurs courbes de degré élevé, ce qui engendre un temps de calcul long.

Dans le cadre de nos travaux, nous avons opté pour une approche mixte, qui utilise à la fois les contours actifs et les modèles paramétriques.
Nous verrons dans les chapitres suivants, que l'utilisation d'un type de snake, appelé « jumping snake » permet de pouvoir initialiser les snakes relativement loin des contours des lèvres et d'obtenir rapidement la position de plusieurs points clefs sur les contours extérieur et intérieur. Ensuite, ces différents points clefs permettent de positionner des modèles paramétriques composés de plusieurs cubiques qui sont suffisamment flexibles pour représenter des variations importantes des formes labiales. La phase d'optimisation est, quant à elle, rapide du fait de la connaissance de nombreux points fournis par la convergence des jumping snakes.

CHAPITRE 4

Modèles paramétriques des lèvres et traitements préliminaires

Pour obtenir un algorithme de segmentation des contours des lèvres robuste et précis, nous combinons contours actifs et modèles paramétriques, pour exploiter les avantages de ces 2 méthodes.

Dans ce chapitre, nous présentons les modèles déformables choisis (jumping snakes et modèles paramétriques), les gradients développés pour accentuer le contour des lèvres et les bases de données utilisées.

La partie 4.1 décrit les modèles paramétriques que nous avons défini pour représenter les contours extérieur et intérieur des lèvres. L'utilisation de courbes cubiques permet d'obtenir un modèle de contours complet suffisamment flexible pour reproduire un panel très large de formes de bouche.

Dans la partie 4.2, nous présentons l'algorithme jumping snake qui est le type de contour actif que nous avons choisi pour notre étude. Le jumping snake possède les avantages de réduire son initialisation au positionnement d'un germe et de permettre un réglage simple de ses paramètres. Ceci est intéressant, dans notre cas, pour obtenir un algorithme automatique et utilisable pour des images variées (conditions d'illumination et tailles de bouche différentes).

Pour la convergence des snakes et l'optimisation des modèles paramétriques, nous utilisons tout au long de notre étude une méthode de maximisation de flux moyen de gradient. Les différents gradients développés pour accentuer chaque partie des lèvres sont exposés dans la partie 4.3.

La partie 4.4 présente les bases d'images que nous avons utilisées dans cette étude pour développer les algorithmes de segmentation et de suivi des contours labiaux externe et interne, ainsi que pour évaluer les performances des méthodes proposées.

4.1. Les modèles paramétriques

Dans la section 3.2.1, nous avons montré que de nombreuses études ont proposé différents modèles paramétriques pour caractériser les contours extérieur et intérieur des lèvres. Nous avons vu également que les modèles les plus courants constitués de paraboles sont trop simplistes pour représenter précisément les contours des lèvres. Nous avons choisi d'utiliser le modèle extérieur introduit par Eveno *et al.* [Eveno 2004] et nous avons proposé 2 modèles paramétriques intérieurs différents selon que la bouche est ouverte ou fermée.

4.1.1. Modèle paramétrique pour le contour extérieur

Le modèle paramétrique utilisé pour le contour extérieur a été proposé par Eveno *et al.* [Eveno 2004]. Il est composé d'une ligne brisée et de 4 cubiques reliées par 6 points clefs P_1 à P_6 (cf. Fig. 4.01).

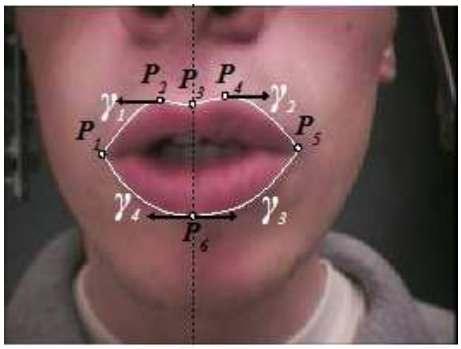


Fig. 4.01. Modèle paramétrique externe.

P_1 et P_5 sont les coins extérieurs de la bouche, appelés également commissures des lèvres. P_2, P_3 et P_4 définissent l'arc de Cupidon, qui est la forme en « V » visible au milieu du contour supérieur de la bouche. P_6 est le point bas du contour extérieur se trouvant sur la verticale passant par P_3.

En plus des 2 segments $[P_2P_3]$ et $[P_3P_4]$, 4 courbes cubiques complètent le modèle extérieur :

- γ_1 entre P_1 et P_2 pour le contour extérieur supérieur gauche,
- γ_2 entre P_4 et P_5 pour le contour extérieur supérieur droit,
- γ_3 entre P_6 et P_5 pour le contour extérieur inférieur droit,
- γ_4 entre P_1 et P_6 pour le contour extérieur inférieur gauche.

Ainsi, le contour extérieur supérieur est défini par les courbes γ_1, γ_2, et par la ligne brisée $[P_2P_3P_4]$, alors que le contour extérieur inférieur est défini par les courbes γ_3 et γ_4.

En prenant en considération la forme particulière du contour extérieur, Eveno *et al.* introduisent les contraintes suivantes :

- une dérivée nulle pour la cubique γ_1 au point P_2,
- une dérivée nulle pour la cubique γ_2 au point P_4,
- une dérivée nulle pour les cubiques γ_3 et γ_4 au point P_6.

En outre, ces contraintes sur les dérivées des courbes cubiques permettent de réduire le nombre de paramètres à estimer lors de la phase d'optimisation du modèle.

Ce modèle a l'avantage de fournir une description précise du contour extérieur des lèvres, sans être trop complexe (connaissant pour chaque cubique les deux points d'extrémité et une valeur de dérivée,

il ne reste plus qu'un seul paramètre à déterminer). De plus, ce modèle est assez flexible pour représenter des formes très variées de bouche. Entre autre, aucune symétrie du modèle n'est imposée.

La figure 4.02 montre des exemples de modèles du contour extérieur des lèvres proposés dans la littérature (cf. Chapitre 2, pour plus de détails). Les modèles à base de paraboles (cf. Fig. 4.02.a et 4.02.b) ou de quartiques (cf. Fig. 4.02.c) ne sont pas appropriés pour des bouches asymétriques du fait de la simplicité du modèle (modèles avec paraboles) ou de la symétrie imposée pour réduire la complexité du modèle (modèles avec quartiques). Le modèle proposé par Eveno (cf. Fig. 4.02.d) permet de décrire précisément le contour extérieur, même dans le cas de grimaces.

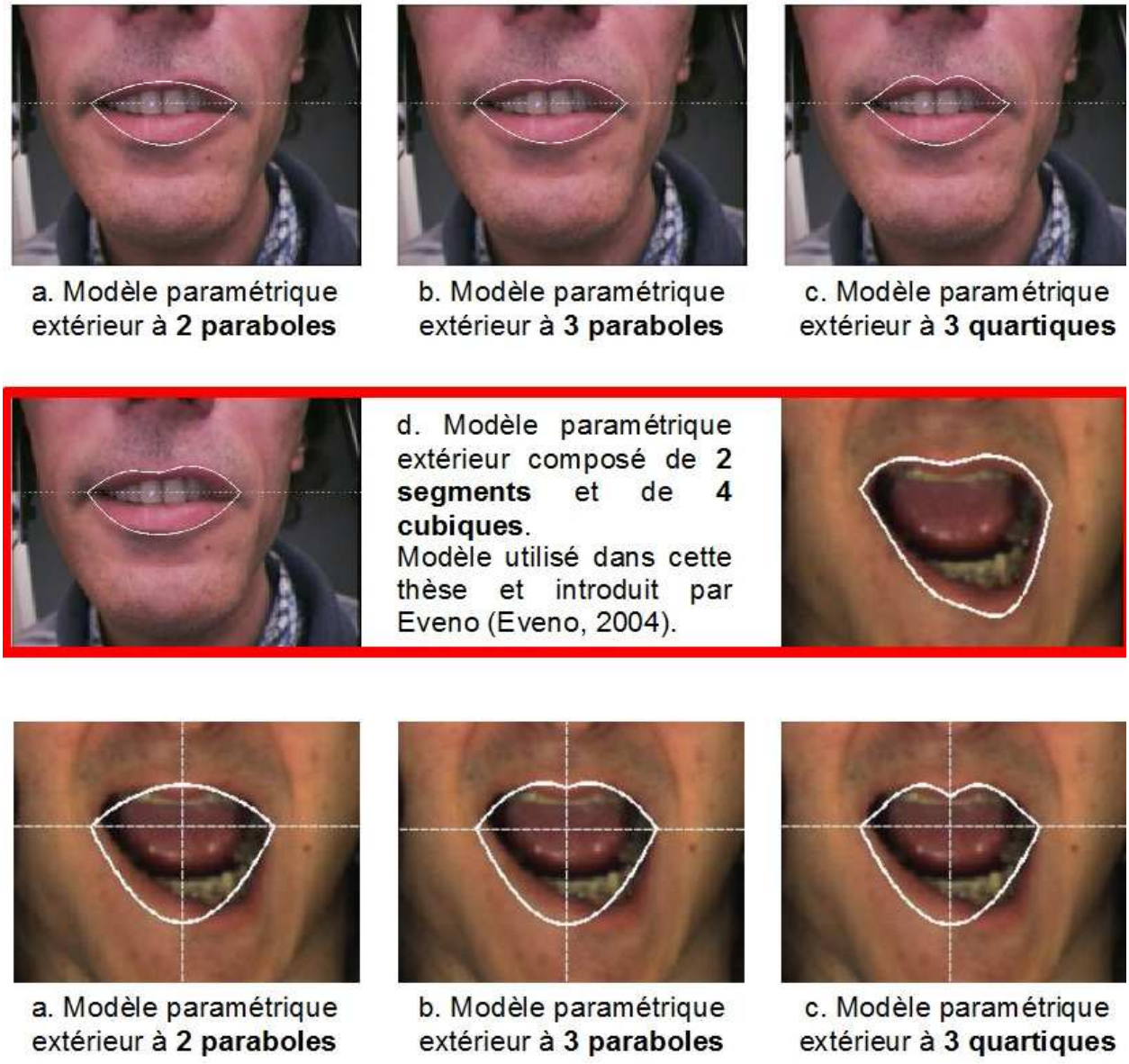


Fig. 4.02. Modèles externes proposés dans la littérature.

4.1.2. Modèle paramétrique pour le contour intérieur

4.1.2.a. Choix des courbes du modèle intérieur

Le choix du modèle paramétrique est important car c'est la variation des paramètres des courbes composant le modèle qui va permettre d'extraire les contours labiaux. Nous avons vu dans la section 3.2 que, dans la littérature, le modèle intérieur est plus simple que le modèle extérieur et que généralement, il est défini par des paraboles.

Une des applications visée par nos travaux est la lecture labiale. Or, les paramètres labiaux utilisés pour la reconnaissance de la parole sont souvent calculés à partir du contour intérieur (cf. Section 1.3). Dans le cadre de cette thèse, les paramètres labiaux à estimer pour le projet TELMA sont les hauteur, largeur et aire de l'intérieur de la bouche, ainsi que les deux valeurs de pincement (cf. Section 1.2.3). En conséquence, le contour intérieur doit être détecté très précisément.

Lors de notre étude, nous avons constaté que l'utilisation de seulement deux paraboles (comme choisi dans la majorité des travaux que nous avons trouvés) pour représenter le contour intérieur (une pour le contour supérieur et une autre pour le contour inférieur) n'était pas suffisant du fait de formes de bouche souvent asymétriques. Le fait d'utiliser quatre paraboles permet de lever la contrainte de symétrie. Nous avons en outre choisi d'utiliser des courbes de degré supérieur car cela permet de segmenter le contour intérieur de manière plus précise. Par exemple, il est possible, suivant les sujets, que l'arc de Cupidon (forme en « V ») soit également visible sur le contour intérieur supérieur (cf. Fig. 4.03). Dans ce cas, et à l'instar du cas extérieur, il est nécessaire d'utiliser des cubiques ou des quartiques (cf. Section 3.2). De plus, il faut être capable de segmenter les contours pour chacun des visèmes (unité élémentaire visuelle, équivalent du phonème du domaine acoustique) possibles. Or, suivant le son prononcé, la bouche prend des formes plus ou moins complexe. Par exemple, lorsque la bouche est arrondie (son /o/), la déformation des lèvres ne peut être représentée par des paraboles.

Fig. 4.03. Arc de Cupidon visible sur le contour extérieur supérieur des lèvres [Martinez, 1998].

Nous avons donc opté pour des courbes d'ordre 3, ce qui, en plus, permet de garder une certaine cohésion avec le modèle paramétrique extérieur composé de cubiques.

Nous avons défini 2 modèles pour modéliser le contour intérieur des lèvres, l'un pour une bouche ouverte et l'autre pour une bouche fermée. Ce choix a été motivé par le fait que ces 2 cas sont très différents. Lorsque la bouche est ouverte, la frontière entre les lèvres et l'intérieur de la bouche peut prendre 4 configurations possibles : Lèvre/Dent, Lèvre/Gencive, Lèvre/Langue ou Lèvre/Cavité orale. Alors que lorsque la bouche est fermée, le contour intérieur a toujours le même aspect et peut être vu comme une ligne sombre séparant la lèvre supérieure de la lèvre inférieure.

4.1.2.b. Cas bouche ouverte

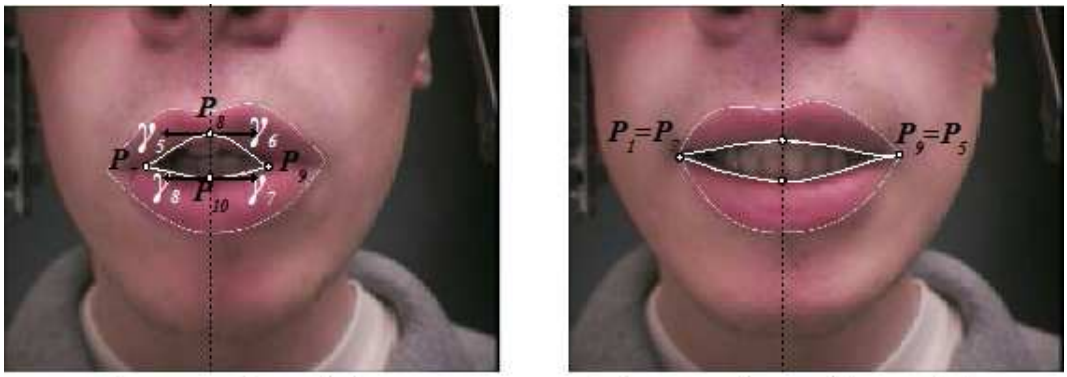
Lorsque la bouche est ouverte, le contour intérieur est défini par 4 cubiques reliées par 4 points clefs (cf. Fig. 4.04). P_8 et P_{10} sont les points milieu du contour intérieur supérieur et inférieur situés sur la verticale passant par P_3. P_3 étant le point milieu de l'arc de Cupidon, la verticale passant par P_3 représente effectivement le milieu de la bouche. P_7 et P_9 sont les commissures internes de la bouche. Initialement [Stillittano 2008], nous avions proposé un modèle où les commissures internes étaient choisies égales aux commissures externes (P_7=P_1 et P_9=P_5). Le choix de différencier les commissures a été motivé par le fait, qu'en cas de mouvements labiaux de type protrusion, les commissures internes et externes diffèrent (cf. Fig. 4.04.a). Pour des applications de lecture labiale, il doit être possible de calculer différents paramètres labiaux comme les largeurs externe et interne de la bouche. Cela permet également une segmentation plus fine lorsque la bouche est légèrement ouverte.

Le modèle du contour intérieur pour une bouche ouverte est complété par 4 cubiques :

- γ_5 entre P_7 et P_8 pour le contour intérieur supérieur gauche,
- γ_6 entre P_8 et P_9 pour le contour intérieur supérieur droit,
- γ_7 entre P_{10} et P_9 pour le contour intérieur inférieur droit,
- γ_8 entre P_7 et P_{10} pour le contour intérieur inférieur gauche.

Le contour interne supérieur est défini par les 2 cubiques γ_5 et γ_6. Le contour interne inférieur est défini par les 2 cubiques γ_7 et γ_8. De la même façon que pour le modèle extérieur, on fait les hypothèses supplémentaires suivantes :

- une dérivée nulle pour les cubiques γ_5 et γ_6 au point P_8,
- une dérivée nulle pour les cubiques γ_7 et γ_8 au point P_{10}.



a. Avec commissures internes

b. Avec commissures internes égales aux commissures externes

Fig. 4.04. Modèle paramétrique interne : cas bouche ouverte.

Ce modèle permet une représentation précise et flexible du contour intérieur des lèvres quand la bouche est ouverte. Chaque cubique étant défini par 2 points et une valeur de dérivée, il reste un seul paramètre par cubique à estimer.

Suivant l'ouverture de la bouche, les commissures internes P_7 et P_9 se trouvent à l'intérieur de la bouche (cf. Fig. 4.04.a), ou elles peuvent être confondues avec les commissures externes P_1 et P_5 du modèle paramétrique extérieur (cf. Fig. 4.04.b).

La figure 4.05 montre des exemples de modèles du contour intérieur des lèvres pour des bouches ouvertes. Les modèles à base de paraboles (cf. Fig. 4.05.a) ne donnent pas autant de précision que le modèle proposé dans ce travail (cf. Fig. 4.05.b). De plus, lorsque la bouche est trop asymétrique (exemple de la grimace), le modèle à base de paraboles ne peut suivre les contours, car il est trop limité par les contraintes de symétrie. En revanche, le modèle à base de cubiques permet de modéliser les contours intérieurs même lorsque les déformations sont importantes.

a. Modèle paramétrique intérieur à **2 paraboles**

b. Modèle paramétrique intérieur à **4 cubiques** utilisé dans cette thèse

Fig. 4.05. Comparaison entre le modèle proposé et le modèle classique de la littérature.

4.1.2.c. Cas bouche fermée

Lorsque la bouche est fermée, le modèle du contour intérieur est composé de 2 cubiques reliées par 1 seul point clef P_{11} (cf. Fig. 4.06). P_{11} est le point milieu du contour intérieur. Pour une bouche fermée, nous considérons que les commissures internes sont les mêmes que les commissures externes (P_7=P_1 et P_9=P_5). Le choix effectué permet une transition simple modèle bouche ouverte/bouche fermée en cas de segmentation d'une séquence vidéo.

Le modèle du contour intérieur pour une bouche fermée est complété par 2 cubiques :

- γ_9 entre P_1 et P_{11} pour le contour intérieur gauche,
- γ_{10} entre P_{11} et P_5 pour le contour intérieur droit.

Comme hypothèses supplémentaires, nous imposons une dérivée nulle pour les cubiques γ_9 et γ_{10} au point P_{11}.

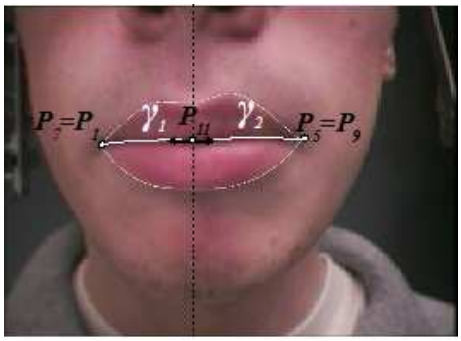


Fig. 4.06. Modèle paramétrique interne : cas bouche fermée.

La figure 4.07 montre des exemples de modèles du contour intérieur des lèvres pour des bouches fermées. Le modèle courant de la littérature composé de 2 paraboles (cf. Fig. 4.07.b) est comparé au modèle proposé (cf. Fig. 4.07.c). Du fait de sa plus grande complexité (courbes d'ordre 3), le modèle proposé offre une plus grande précision. Dans le cas d'une bouche fermée, le contour intérieur n'est qu'une ligne sombre séparant les lèvres supérieure et inférieure, et un modèle composé de courbes d'ordre 2 est suffisant dans la majorité des cas. Mais nous avons proposé un modèle à 2 cubiques pour permettre une transition simple entre le modèle interne et le modèle externe (composé de 4 cubiques) dans le cas d'un suivi des contours labiaux dans une vidéo.

Fig. 4.07. Comparaison entre le modèle courant de la littérature et le modèle proposé.

4.2. L'algorithme jumping snake

Dans le chapitre 3, nous avons vu que l'initialisation et le réglage des paramètres des contours actifs sont des facteurs déterminants pour obtenir une bonne segmentation, tout en étant les plus difficiles à choisir. En effet, l'initialisation doit être suffisamment près du contour à détecter pour ne pas que le contour actif soit attiré par un autre minimum local. Aussi, le réglage des paramètres se fait souvent expérimentalement et un jeu de paramètres est rarement réutilisable pour d'autres types d'image. Le jumping snake introduit par Eveno *et al.* [Eveno, 2003] est un contour actif qui converge en une succession de phases de saut et de croissance. La phase d'initialisation est réduite au positionnement d'un germe, qui peut être relativement loin du contour final, et le réglage des paramètres est simple et intuitif.

La figure 4.08 montre le schéma global du fonctionnement de l'algorithme, illustré dans le cas de la segmentation du contour extérieur supérieur des lèvres. Le jumping snake est initialisé par un germe S^0, qui peut être choisi loin du contour (une discussion sur le choix de la position du germe et sur les réglages des paramètres est présentée par la suite). Ensuite, pendant la phase de croissance, des points sont ajoutés à gauche et à droite du germe pour former une chaîne de points. Finalement, le germe saute vers une

nouvelle position qui est plus proche du contour recherché. Le processus de croissance et de saut est répété jusqu'à ce que l'amplitude du saut soit inférieure à un certain seuil.

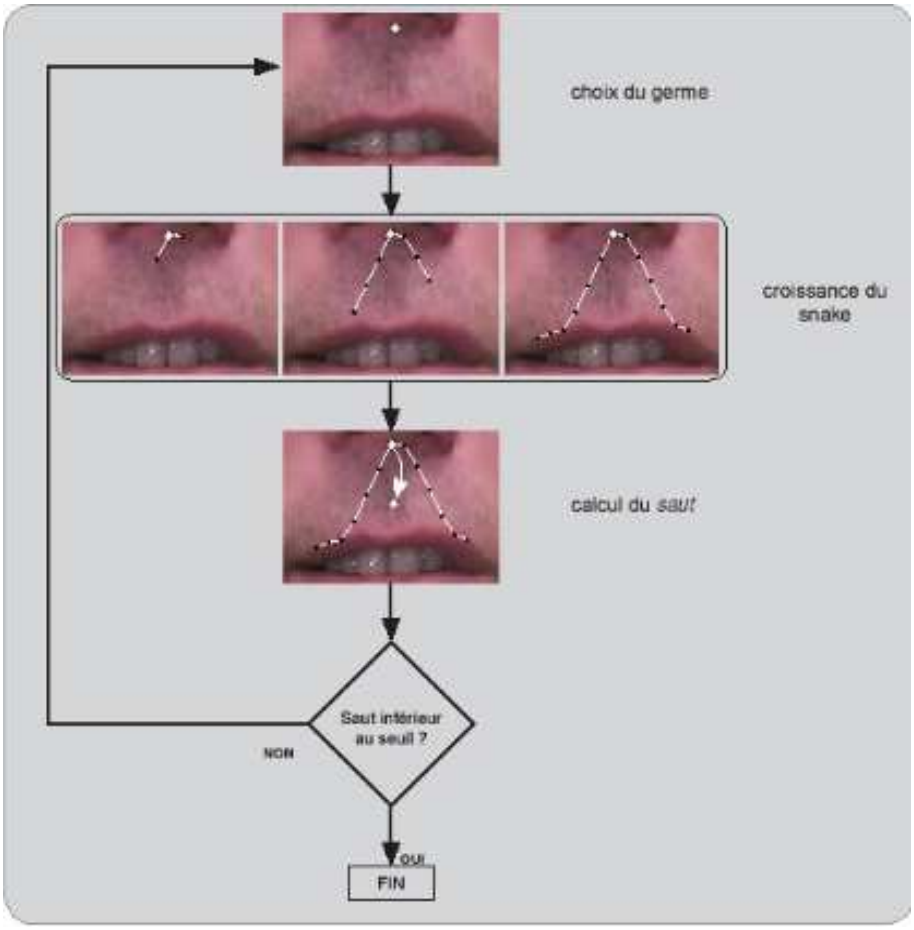


Fig. 4.08. L'algorithme jumping snake [Eveno, 2003].

4.2.1. Paramétrage du jumping snake

Dans [Eveno, 2003], Eveno *et al.* présentent le jumping snake comme étant un contour actif avec des paramètres faciles à régler. Les différents paramètres sont soient choisis manuellement (position du germe initial) soient définis expérimentalement.
Dans le cadre de notre travail, il a fallu définir des règles plus rigoureuses pour l'étape de paramétrisation car nous utilisons plusieurs jumping snakes. La position des différents germes et le réglage des paramètres suivant l'image à traiter sont des choix cruciaux pour une bonne segmentation des lèvres.

4.2.1.a. Choix du germe initial S^0

Le fait que l'étape initiale de l'algorithme du jumping snake se réduise au positionnement d'un seul point (le germe), rend l'initialisation moins problématique que pour un contour actif classique. En effet, le déroulement discontinu de la méthode durant la phase de croissance permet de franchir plus facilement des zones bruitées.

Par la suite, nous utilisons l'algorithme du jumping snake pour trouver des points clefs sur les 4 contours des lèvres (contours extérieurs haut et bas, et contours intérieurs haut et bas), en faisant converger un snake pour chacun des 4 cas. Les contraintes pour la position du germe S^0 sont restreintes aux 2 règles suivantes :

1) Le déroulement de l'algorithme fait que l'on obtient autant de points à gauche et à droite du germe initial S^0 (N points) à la fin de la 1ère convergence. Ainsi, il faut que la position horizontale du germe initial soit proche de la colonne passant au milieu de la bouche, pour ne pas obtenir un snake final décalé.
2) S^0 doit se trouver plus proche du contour recherché que d'un autre minimum local.

Ainsi, compte tenu de la forme des contours des lèvres, par la suite, nous choisissons les germes de la manière suivante :

- Pour le contour extérieur haut, le germe doit se trouver au dessus de la bouche et il doit être plus proche de la bouche que du nez.
- Pour le contour extérieur bas, le germe doit se trouver en dessous de la bouche et il doit être plus proche de la bouche que du menton.
- Pour le contour intérieur haut, le germe doit se trouver en dessus du contour et il doit être plus proche du contour intérieur haut que du contour extérieur haut.
- Pour le contour intérieur bas, le germe doit se trouver au dessous du contour et il doit être plus proche du contour intérieur bas que du contour extérieur bas.

La figure 4.09 montre une illustration des zones d'initialisation privilégiées pour les 4 germes.

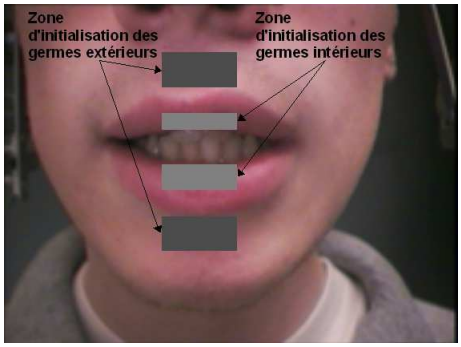


Fig. 4.09. Zone d'initialisation des germes.

4.2.1.b. Réglages des paramètres

Le réglages des paramètres d'un jumping snake comporte 3 étapes : le choix de N qui donne le nombre total de points du snake ($2N$+1), le choix de l'espacement horizontal Δ entre chaque point du snake (cf. Fig. 4.10) et le choix des candidats possibles à tester lorsque l'on ajoute un point au snake (nombre et position des candidats).

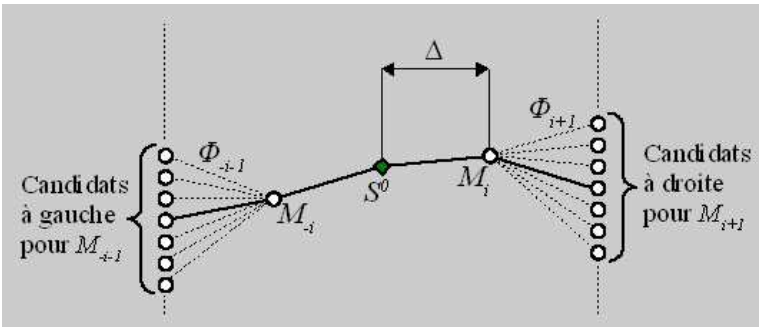


Fig. 4.10. Paramètres du jumping snake.

Réglage de N et Δ :

Dans le contexte de la segmentation des lèvres, nous verrons par la suite que nous utilisons 4 jumping snakes pour obtenir la position des 6 points clefs suivants : P_2, P_3, P_4, P_6, P_8 et P_{10} (cf. Fig. 4.01 et 4.04).

En théorie N pourrait être égal à 1. On aurait ainsi des snakes constitués de 3 points et il serait facile d'associer les points du snake aux différents points clefs que l'on recherche. Cependant, ce nombre doit être plus grand pour 2 raisons :

1) Un snake, avec une valeur de N faible, engendre une zone de convergence restreinte et le résultat final est très proche de la position du germe initial S^0. Or, nous voulons justement utiliser l'algorithme du jumping snake pour permettre une initialisation plus éloignée du contour. Plus N est grand et plus le snake pourra franchir facilement les zones bruitées.
2) Nous verrons dans le chapitre 5 que les jumping snakes servent à trouver la position de tous les points clefs des modèles paramétriques présentés dans la partie 4.1, exceptés la position des commissures (P_1 et P_5) qui demande une approche différente du fait de leur position particulière. Ensuite, les points clefs sont reliés par des courbes cubiques calculées par la méthode des moindres carrés. Pour que la méthode fournissent des résultats satisfaisants, il faut connaître la position de trois points, en plus des deux points clefs extrêmes, pour chacune des cubiques. La figure 4.11 montre la position des trois points supplémentaires fournis par les snakes (points verts) pour calculer les cubiques $\gamma_{i=1 \text{ à } 8}$.
Si nous prenons le cas du contour extérieur supérieur, il faut, au minimum, que le snake comporte neuf points (trois points pour P_2, P_3 et P_4, trois points supplémentaires pour γ_1 et trois points supplémentaires pour γ_2), ce qui correspond à $N=4$.

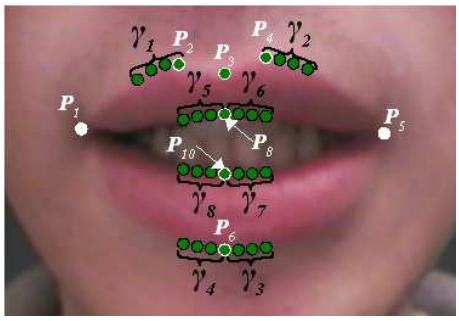


Fig. 4.11. Points utiles pour le calcul des cubiques.

En conclusion, N doit être **supérieur ou égal à 4**. Nous allons voir que le choix final de la valeur de N se fait en fonction de la valeur de Δ et de la largeur de la bouche.

Le paramètre Δ permet de régler la finesse du contour obtenu. Plus Δ est petit et plus les détails du contour sont rendus précisément, mais il faudra une valeur de N importante, et donc une complexité de calcul plus importante, pour avoir un snake suffisamment large. A l'inverse, avec une valeur de Δ grande, le résultat de la segmentation est grossier. Expérimentalement, nous avons constaté que **Δ=6** pixels permet un calcul rapide, tout en ayant un espacement suffisant pour calculer un flux moyen de gradient significatif entre deux points successifs du snake.

La figure 4.12 montre des exemples de convergence de jumping snake pour la segmentation du contour extérieur supérieur pour différentes valeurs de N et Δ, avec N=Δ. Il n'est pas obligatoire de prendre N=Δ, mais c'est le choix que nous avons adopté par la suite, car cela va permettre de déterminer les valeurs de N et Δ automatiquement en fonction de la largeur de la bouche. La largeur de la bouche de

l'image de test est de 85 pixels. La ligne blanche représente la dernière itération.
Nous pouvons remarquer que le temps de calcul (t_{conv}, en seconde et pour une implémentation en Matlab) est étroitement lié aux valeurs de N et Δ. Le premier germe S^0 étant le même pour chaque cas, le nombre d'itérations nécessaires pour obtenir le snake final est différent. Il peut soit diminuer soit augmenter quand N et Δ augmentent, dans la mesure où le germe à l'itération suivante est calculé en fonction du résultat du snake courant.

Ainsi, le choix de N et Δ est un compromis entre vitesse de convergence et finesse du contour final. Dans l'exemple de la figure 4.12, le couple donnant le meilleur compromis est (N=6, Δ=6). Ce qui correspond à un snake de longueur $\Delta*(2N)$=72 pixels, proche des 85 pixels de la largeur de la bouche de l'image de test. Pour la suite de notre étude, nous avons choisi de déterminer N et Δ en fonction de la largeur de la bouche (évaluée à partir de la position des yeux trouvée par l'algorithme C3F, cf. Section 5.1) en imposant que $2N\Delta$ soit environ égal à la largeur de la bouche.

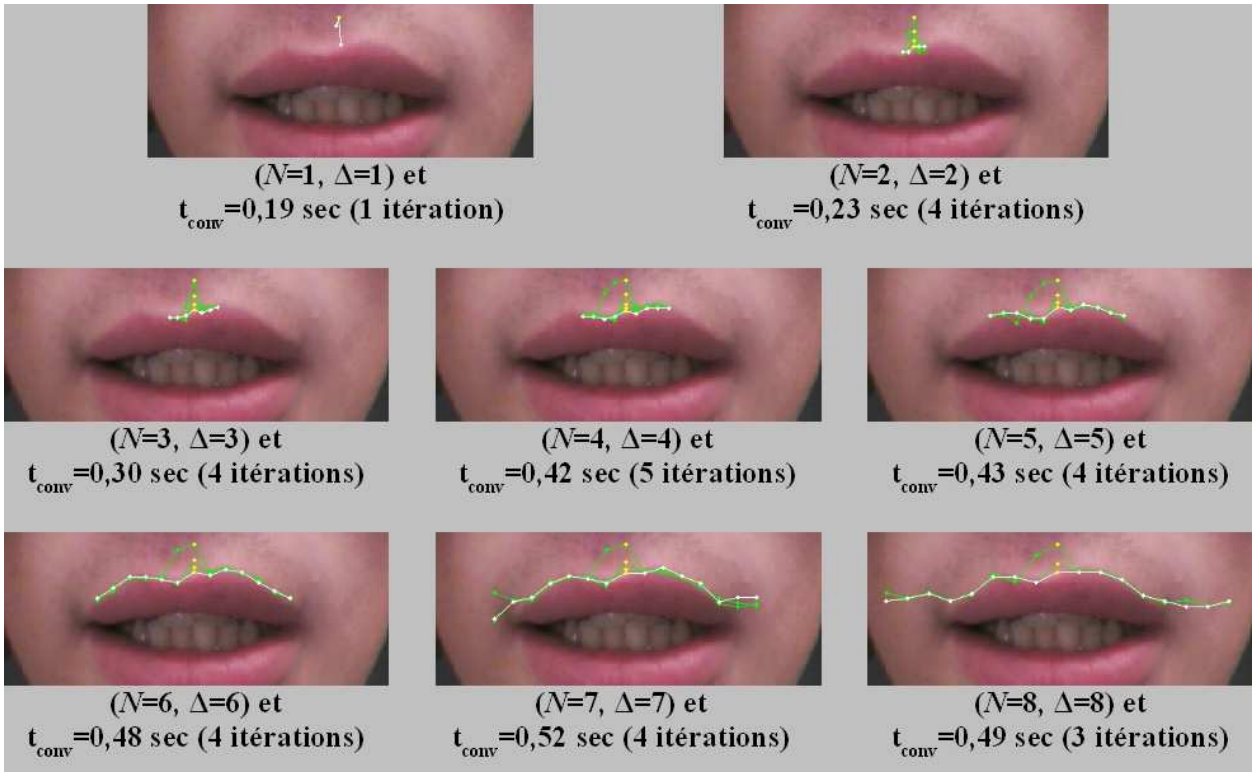


Fig. 4.12. Convergence pour différentes valeurs N et Δ.

Les images sur lesquelles nous avons travaillé se découpent en 2 catégories : des images et des séquences vidéo où la largeur de la bouche va de 80 à 120 pixels, et une base de séquences vidéo où la largeur moyenne de la bouche est de 50 pixels.
Compte tenu du fait qu'un snake aura une taille de $\Delta*(2N)$ pixels, nous avons choisi 3 cas de figure et 3 couples (N, Δ) différents suivant la catégorie d'image testée.

- (N=5, Δ=5) pour les séquences vidéo où la bouche fait 50 pixels de largeur en moyenne (ce qui donne un snake de longueur $\Delta*(2N)$=50 pixels),
- (N=6, Δ=6) pour les images où la largeur de bouche est comprise entre 80 et 100 pixels (ce qui donne un snake de longueur $\Delta*(2N)$=72 pixels),
- (N=7, Δ=7) pour les images où la largeur de bouche est comprise entre 100 et 120 pixels (ce qui donne un snake de longueur $\Delta*(2N)$=98 pixels).

Il est possible que le snake se retrouve au centre de la bouche sans atteindre ses extrémités (si la longueur du snake est plus petite que la largeur de la bouche), mais cela ne pose pas de problème, car comme nous avons vu avec la figure 4.11, les points nécessaires au calcul des cubiques donnés par les 4 snakes sont positionnés au milieu de la bouche.
Avec ces trois choix de couple (N, Δ), nous avons donc un algorithme qui peut être utilisé pour des images où la largeur de bouche est comprise entre 50 et 120 pixels. Pour des images où la bouche serait plus petite que 50 pixels ou plus grande que 120 pixels, il suffit de redimensionner l'image pour obtenir une largeur de bouche correspondant à l'un des 2 cas extrêmes.
La figure 4.13 montre des exemples de convergence des 4 jumping snakes pour différentes tailles de bouche. Pour une meilleure visibilité, les échelles ne sont pas respectées.

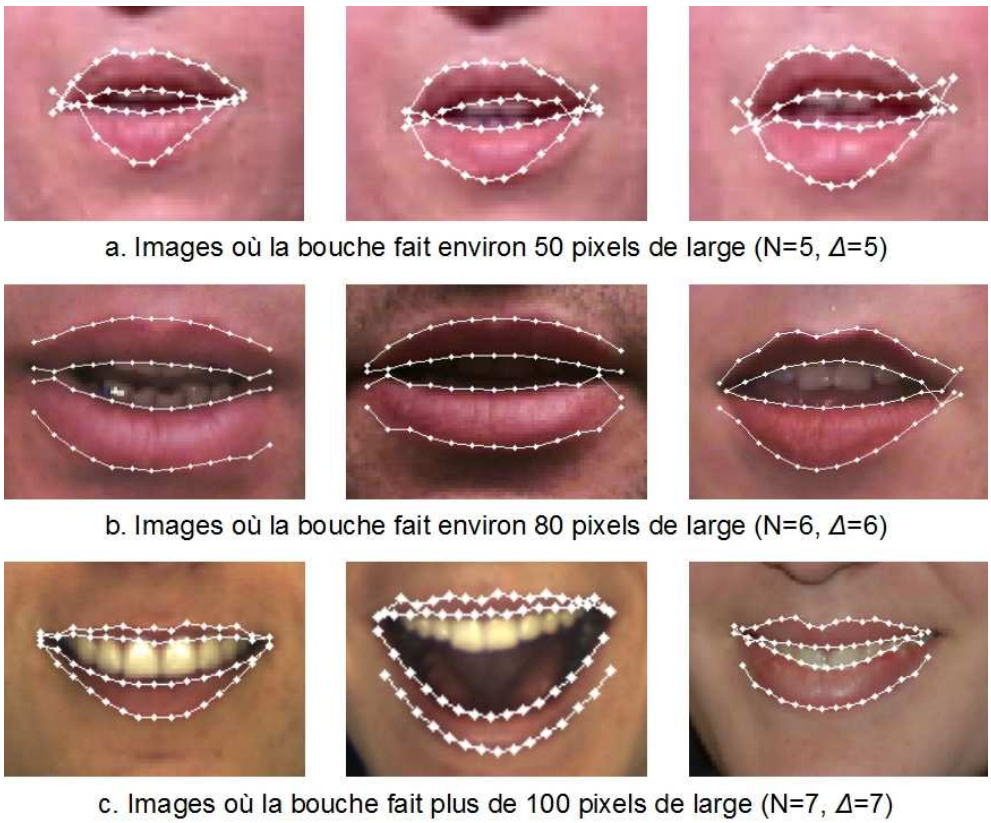


Fig. 4.13. Exemples de convergence des snakes pour différentes tailles de bouche.

Choix des candidats possibles :

Lorsqu'un point est ajouté au snake pendant la phase de croissance, plusieurs points candidats sont testés (positionnés sur la colonne distante de ±Δ, cf. Fig. 4.10). Considérant que le point se trouvant sur la même ligne que le point terminal courant du snake est toujours testé, il faut déterminer le nombre N_{up} et N_{low} de points testés au dessus et au dessous de la ligne. A partir de là, il faut différencier 2 cas possibles : soit nous voulons que le snake se propage au dessus du germe, soit nous voulons qu'il se propage en dessous. Dans le premier cas, il faut $N_{up} > N_{low}$ et dans le second cas, il faut $N_{up} < N_{low}$. Les valeurs de N_{up} et N_{low} sont fixées expérimentalement. Prenons comme exemple le cas du jumping snake utilisé pour le contour extérieur supérieur. Le germe se situe au dessus de la bouche, entre le nez et la lèvre supérieure, et le snake doit se propoager vers le bas ($N_{up} < N_{low}$). Il faut que N_{low} soit suffisamment grand pour aller jusqu'à la bouche, mais la valeur ne doit pas être trop élevée pour éviter de passer par dessus le contour. Il faut que N_{up} soit plus petit que N_{low}, mais la valeur ne doit pas être trop faible, pour pouvoir compenser la trajectoire descendante et freiner le snake au niveau du contour. Ces valeurs sont

choisies en fonction de la résolution de l'image. En ce qui concerne nos images, nous avons choisi 2 valeurs possibles : 3 ou 10 pixels. N_{up} = 3 et N_{low} = 10 pour le cas descendant et N_{up} = 10 et N_{low} = 3 pour le cas montant. Ceci permet d'avoir de bons résultats de convergence pour les images quelque soit la largeur de la bouche (allant de 50 à 120 pixels).

4.2.1.c. Utilisation des jumping snakes pour localiser des points clefs

Les contours actifs sont largement utilisés pour la segmentation des lèvres. Cependant le problème du positionnement de la courbe initiale est problématique car celle-ci doit être proche du contour recherché pour ne pas être attirée par un contour parasite. De plus, les paramètres des contours actifs sont souvent difficiles à régler. L'algorithme du jumping snake permet de résoudre en partie ces 2 problèmes. Dans notre cas, cet algorithme sert avant tout à trouver des points clefs pour initialiser nos différents modèles paramétriques. Quatre différents jumping snakes sont pris en compte :

- un snake extérieur supérieur pour le contour extérieur supérieur des lèvres,
- un snake extérieur inférieur pour le contour extérieur inférieur des lèvres,
- un snake intérieur supérieur pour le contour intérieur supérieur des lèvres,
- un snake intérieur inférieur pour le contour intérieur inférieur des lèvres.

Pour faire un parallèle avec les contours actifs classiques [Kass, 1987] présentés dans la section 2.4.1, l'énergie interne du jumping snake, permettant de régulariser la courbe pendant les déformations, est représentée par le paramètre Δ. En effet, Δ impose des contraintes sur la forme du snake en forçant les points à être espacés du même nombre de colonne dans l'image. L'énergie externe est choisie comme étant un flux moyen de gradient Φ entre les points du snake (cf. Fig. 4.10). La section suivante décrit les 4 gradients utilisés pour la convergence de nos 4 jumping snakes.

4.3. Gradients développés pour accentuer le contour des lèvres

Cette section présente les 4 gradients que nous avons construits pour accentuer le contour des lèvres. Ces gradients seront utilisés d'une part pour la convergence de différents jumping snakes et d'autre part, pour l'optimisation des modèles paramétriques que nous avons proposés.

4.3.1. Les plans couleurs

Nous utilisons des gradients calculés grâce à plusieurs composantes de différents espaces couleurs. Nous employons les espaces couleur *RGB* et *Luv*, ainsi que les pseudo-teintes *H1* et *H2*. *H1* correspond au canal *U* (cf. Section 2.1.2.b, [Liévin, 2004]), et *H2* à la pseudo-teinte $\hat{H}$ (cf. Section 2.1.2.a, [Poggio, 1998]).

Toutes les composantes sont normalisées entre 0 et 1, si cela n'a pas été déjà fait par construction, pour avoir une meilleure dynamique et combiner des valeurs équivalentes. Pour trouver la valeur normalisée $C_N(x,y)$ du pixel (x,y), nous utilisons les valeurs maximale $max(C)$ et minimale $min(C)$ trouvées dans toute l'image et la valeur de la composante non normalisée $C(x,y)$:

$$C_N(x,y) = \frac{C(x,y) - min(C)}{max(C) - min(C)} \tag{4.01}$$

Par la suite, nous utiliserons plusieurs composantes couleurs ou de luminance normalisées pour créer des gradients accentuant les différents contours des lèvres. Nous pouvons citer $H1_N$ (Fig. 4.14.b.) et $H2_N$ (Fig. 4.14.c) qui augmentent le contraste entre les lèvres et la peau, L_N (Fig. 4.14.d), la luminance qui a des valeurs plus fortes au dessus des lèvres et u_N (Fig. 4.14.e) qui a des valeurs faibles pour les dents (les 2 dernières composantes viennent de l'espace *Luv*). L'image de la figure 4.14.a provient de la base d'images AR [Martinez, 1998].

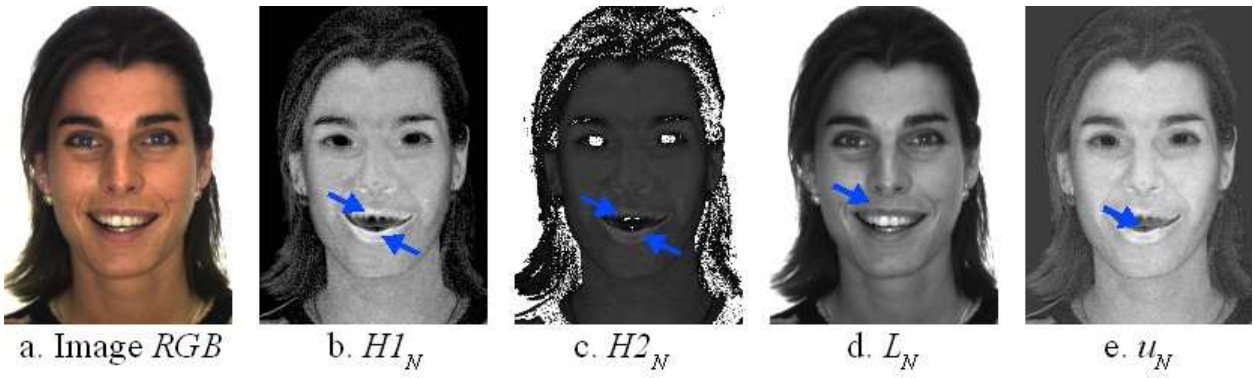


Fig. 4.14. Composantes utilisées pour créer les gradients des contours des lèvres.

4.3.2. Les gradients G_1 et G_2 pour le contour extérieur

4.3.2.a. Le gradient G_1 pour le contour extérieur supérieur

Pour accentuer le contour extérieur supérieur, nous utilisons le gradient hybride, noté G_1 ici, proposé dans [Eveno, 2003] et calculé avec l'équation suivante :

$$G_1(x,y) = \nabla[+H1_N(x,y) - L_N(x,y)] \tag{4.02}$$

∇ est l'opérateur gradient. $H1_N(x,y)$ est la pseudo-teinte *H1* normalisée et $L_N(x,y)$ est la luminance normalisée, issue de l'espace couleur *Luv*, du pixel (x,y).

Il est à noter qu'au lieu d'utiliser la pseudo-teinte *H2* comme dans la formule originale du gradient hybride (cf. Section 2.2.2), nous utilisons la pseudo-teinte *H1*. En effet, notre travail sur la construction des gradients a montré que le gradient *H1* est plus spécifique au contour extérieur des lèvres que la pseudo-teinte *H2*. Et inversement, *H2* est plus efficace pour accentuer le contour intérieur que *H1* (cf. gradient G_3, eq. 4.04). Ceci s'explique par le fait que le contraste entre les lèvres et la peau est plus accentué avec *H1* qu'avec *H2* (cf. Fig. 4.14.b et c). En revanche, les lèvres et l'intérieur de la bouche sont mieux différenciés avec *H2* qu'avec *H1*, notamment lorsque l'intérieur est sombre, ce qui est généralement le cas.

Le gradient G_1 est utilisé pour le contour extérieur supérieur car la pseudo-teinte *H1* augmente le contraste entre la lèvre supérieure et la peau située au dessus de la bouche (cf. Fig. 4.14.b). Le signe + de l'équation 4.02 vient du fait que les valeurs de *H1* sont plus faibles pour la peau que pour la lèvre supérieure (située en dessous). De plus, généralement, la lumière vient d'en haut. Ainsi, la frontière supérieure de la bouche est une zone de forte luminance (cf. Fig. 4.14.d), alors que le haut de la lèvre supérieure est plus sombre (d'où le signe - devant la luminance dans l'équation 4.02).

La figure 4.15 montre, sur 3 exemples d'images *RGB*, la combinaison des plans avant le calcul du gradient G_1 ($H1_N$ - L_N, cf. Fig. 4.15.b) et le gradient G_1 (cf. Fig. 4.15.c) accentuant le contour extérieur supérieur des lèvres.

Fig. 4.15. Exemples d'accentuation du contour extérieur supérieur avec le gradient G_1.

Sur ces 3 exemples, nous pouvons observer que le gradient G_1, en plus d'accentuer le contour supérieur, peut accentuer d'autres contours de la bouche. Ce même phénomène est visible pour les gradients G_2, G_3 et G_4. Ceci s'explique simplement par le fait que pour créer nos gradients, nous utilisons des combinaisons (somme ou produit) de plusieurs composantes qui peuvent avoir des valeurs fortes ou faibles pour différentes parties de la bouche (lèvres, dents, langue, gencive ou cavité orale).

4.3.2.b. Le gradient G_2 pour le contour extérieur inférieur

Le gradient G_2 utilise également la pseudo-teinte *H1*, mais aussi les composantes de l'espace *RGB*.

$$G_2(x,y)=\nabla\left[-H1_N(x,y)-\left\{R_N(x,y)-G_N(x,y)+B_N(x,y)\right\}\right] \qquad (4.03)$$

∇ est l'opérateur gradient. $H1_N(x,y)$ est la pseudo-teinte *H1* normalisée, $R_N(x,y)$, $G_N(x,y)$ et $B_N(x,y)$ sont les composantes normalisées de l'espace couleur *RGB* du pixel (x,y).

De la même façon, *H1* est utilisée pour sa capacité à faire ressortir les lèvres par rapport à la peau (cf. Fig. 4.14.b). Le signe + de l'équation 4.03 devant *H1* vient du fait que les valeurs de *H1* sont plus fortes pour la lèvre inférieure que pour la peau qui se trouve en dessous. La somme *R-G+B* est une combinaison qui permet d'accentuer les lèvres, d'où le même signe - devant cette somme. Une combinaison semblable est utilisée dans [Jian, 2006].

La figure 4.16 montre, sur 3 exemples d'images *RGB*, la combinaison des plans avant le calcul du gradient G_2 ($-H1_N-[R_N-G_N+B_N]$, cf. Fig. 4.16.b) et le gradient G_2 (cf. Fig. 4.16.c) accentuant le contour extérieur inférieur des lèvres.

Fig. 4.16. Exemples d'accentuation du contour extérieur inférieur avec le gradient G2.

4.3.3. Les gradients G_3 et G_4 pour le contour intérieur

Pour le contour intérieur, la construction de gradient est un challenge plus difficile. En effet, le contour extérieur est toujours une frontière entre des lèvres et de la peau. Le contour extérieur supérieur sépare la peau, située entre le nez et la bouche, et la lèvre supérieure. Le contour extérieur inférieur sépare la peau, située sous la bouche, et la lèvre inférieure. En revanche, pour le contour intérieur, la frontière se situe entre des lèvres et une des 4 possibilités suivantes : dents, gencives, langue ou cavité orale (cf. Fig 4.17). De plus, il faut que le contour soit prononcé pour tous les cas possibles, car durant une conversation, la variation d'apparence de l'intérieur de la bouche est non linéaire et nous pouvons avoir des transitions brutales entre ces configurations. Après expérimentation sur plusieurs centaines d'images, il s'est avéré qu'il était impossible de trouver une composante d'un espace couleur permettant d'obtenir un gradient efficace pour toutes ces configurations. Ainsi, nous avons construit des gradients intérieurs qui sont la combinaison de plusieurs composantes différentes appropriées dans un des cas possibles.

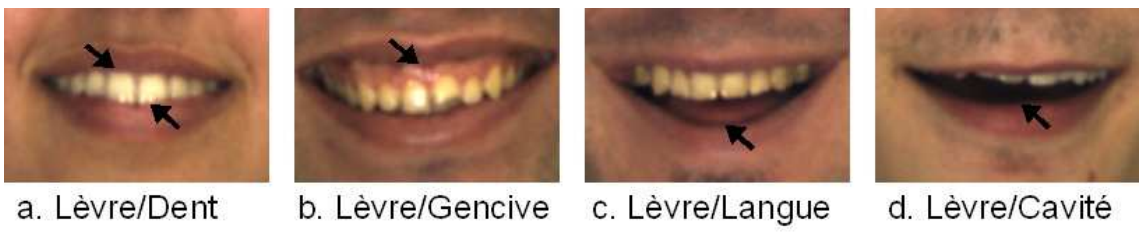


Fig. 4.17. Exemples d'apparences de l'intérieur de la bouche.

4.3.3.a. Le gradient G_3 pour le contour intérieur supérieur

Nous calculons le gradient G_3, intensifiant le contour intérieur supérieur, de la manière suivante :

$$G_3(x,y) = \nabla[+R_N(x,y) - H2_N(x,y) - u_N(x,y) - L_N(x,y)] \qquad (4.04)$$

∇ est l'opérateur gradient. R_N(x,y) est la composante normalisée issue de l'espace couleur *RGB*, $H2_N(x,y)$ est la pseudo-teinte *H2* normalisée, $u_N(x,y)$ et $L_N(x,y)$ sont les composantes normalisées issues de l'espace couleur *Luv* du pixel (x,y).

Le gradient G_3 a été choisi considérant que :

- la composante *R* peut être plus faible pour les pixels de la lèvre supérieure que pour l'intérieur de la bouche (d'où le signe + de l'équation 4.04),
- les valeurs de la pseudo-teinte *H2* (cf. Fig. 4.14.c) sont plus fortes pour les pixels de la lèvre supérieure que pour l'intérieur de la bouche (d'où le signe -),
- la composante *u* est plus grande pour les pixels de la lèvre supérieure que pour des pixels « dent » situés en dessous. En effet *u* est proche de 0 pour les dents (cf. Fig. 4.14.e). (d'où le signe – devant la composante *u*),
- la luminance *L* a, généralement, des valeurs plus grandes pour les pixels de la lèvre supérieure que pour l'intérieur de la bouche (d'où le signe -). En général, l'intérieur de la bouche est sombre, excepté en cas de présence de dents (mais cela est comblé par l'apport de la composante *u*).

La figure 4.18 montre, sur 3 exemples d'images *RGB*, la combinaison des plans avant le calcul du gradient G_3 ($R_N - H2_N - L_N - u_N$, cf. Fig. 4.18.b) et le gradient G_3 (cf. Fig. 4.18.c) accentuant le contour intérieur supérieur des lèvres.

Fig. 4.18. Exemples d'accentuation du contour intérieur supérieur avec le gradient G_3.

4.3.3.b. Le gradient G_4 pour le contour intérieur inférieur

Pour le gradient G_4, nous utilisons un produit des composantes *u* et *L*.

$$G_4(x,y)=\nabla[u_N(x,y)*L_N(x,y)] \quad (4.05)$$

∇ est l'opérateur gradient. * est l'opérateur de multiplication élément par élément; les matrices doivent être de la même taille, ce qui est le cas ici, car les matrices sont les valeurs des composantes pour chaque pixel *(x,y)* de l'image. $u_N(x,y)$ et $L_N(x,y)$ sont les composantes normalisées issues de l'espace couleur *Luv* du pixel *(x,y)*.

Le gradient G_4 a été choisi considérant que :

- la composante *u* est plus grande pour des pixels « lèvre » que pour des pixels « dent » (en effet *u* est proche de 0 pour les dents)
- la luminance *L* a, généralement, des valeurs plus grandes pour des pixels « lèvre » que pour l'intérieur de la bouche

La figure 4.19 montre, sur 3 exemples d'images *RGB*, la combinaison des plans avant le calcul du gradient G_4 (u_N*L_N , cf. Fig. 4.19.b) et le gradient G_4 (cf. Fig. 4.19.c) accentuant le contour intérieur inférieur des lèvres.

Fig. 4.19. Exemples d'accentuation du contour intérieur inférieur avec le gradient G_4.

4.3.4. Filtres utilisés pour le calcul des gradients

Dans de nombreuses études, seule la composante horizontale des gradients est utilisée dans le contexte de la segmentation des lèvres, dans la mesure où c'est la composante prédominante, compte tenu de la forme particulière de la bouche (les contours sont essentiellement horizontaux).
En plus de cette composante, nous avons créé des filtres 2D qui permettent de calculer une composante diagonale (la composante verticale des gradients n'est pas utilisée non plus dans notre travail). Ce choix

est motivé par le fait que les contours intérieur et extérieur des lèvres peuvent être assimilés à 2 losanges (cf. Fig. 4.20).

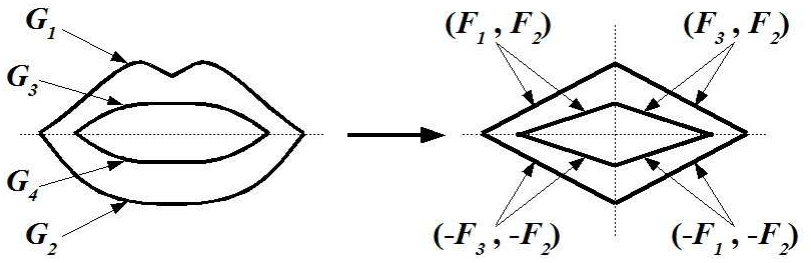
Les 3 filtres utilisés sont les filtres F_1, F_2 et F_3 suivants :

$$F_1=\begin{bmatrix}-2 & -1 & 0\\ -1 & 0 & 1\\ 0 & 1 & 2\end{bmatrix} \quad F_2=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix} \quad F_3=\begin{bmatrix}0 & -1 & -2\\ 1 & 0 & -1\\ 2 & 1 & 0\end{bmatrix} \tag{4.06}$$

Ces filtres (et leurs opposés) permettent d'intensifier la composante diagonale des gradients G_1, G_2, G_3 et G_4 suivant le côté que l'on traite (cf. Fig. 4.20.b.).

- Pour le contour extérieur supérieur gauche, la composante diagonale de G_1 est calculée avec le filtre F_1 et la composante horizontale de G_1 avec le filtre F_2.
- Pour le contour extérieur supérieur droit, la composante diagonale de G_1 est calculée avec le filtre F_3 et la composante horizontale de G_1 avec le filtre F_2.
- Pour le contour extérieur inférieur gauche, la composante diagonale de G_2 est calculée avec le filtre $-F_3$ et la composante horizontale de G_2 avec le filtre $-F_2$.
- Pour le contour extérieur inférieur droit, la composante diagonale de G_2 est calculée avec le filtre $-F_1$ et la composante horizontale de G_2 avec le filtre $-F_2$.
- Pour le contour intérieur supérieur gauche, la composante diagonale de G_3 est calculée avec le filtre F_1 et la composante horizontale de G_3 avec le filtre F_2.
- Pour le contour intérieur supérieur droit, la composante diagonale de G_3 est calculée avec le filtre F_3 et la composante horizontale de G_3 avec le filtre F_2.
- Pour le contour intérieur inférieur gauche, la composante diagonale de G_4 est calculée avec le filtre $-F_3$ et la composante horizontale de G_4 avec le filtre $-F_2$.
- Pour le contour intérieur inférieur droit, la composante diagonale de G_4 est calculée avec le filtre $-F_1$ et la composante horizontale de G_4 avec le filtre $-F_2$.

En conséquence, l'algorithme prend en compte la forme spécifique de la bouche et adapte chaque gradient par rapport au contour traité.



a. Forme réelles des lèvres et gradients utilisés pour chaque élément du contour

b. Forme en losange des lèvres et filtres appliqués selon la partie des lèvres correspondante

Fig. 4.20. Forme des lèvres et filtres 2D associés.

4.4. Les bases d'images

Les bases d'images que nous avons utilisées pour développer et tester notre algorithme sont au nombre de trois :

- des images acquises au sein du laboratoire Gipsa, que nous appellerons base Gipsa,
- la base AR créée par Martinez *et al.* [Martinez, 1998],
- et une base acquise pour le projet TELMA, que nous appellerons base TELMA.

4.4.1. Base Gipsa

Les images, issues de la base de données que nous appelons base Gipsa, sont des images utilisées notamment dans les thèses de Delmas [Delmas, 2000] et de Eveno [Eveno, 2003] sur la segmentation des contours des lèvres. Elles ont été acquises au laboratoire Gipsa à l'aide d'une micro-caméra montée sur un casque fixé sur la tête du locuteur (cf. Fig. 4.21).

Les images sont en RGB (8 bits/couleur/pixel) et elles ont été acquises dans des conditions naturelles (éclairage non uniforme) et sans maquillage particulier. Du fait de la position fixe du casque, le cadrage est constant et centré sur la bouche. La région du visage va du nez jusqu'au cou. Cette base contient 150 images correspondant à six séquences vidéo de 25 images obtenues pour six locuteurs différents. La figure 4.22 montre la première image de chacune des six séquences. La largeur moyenne des bouches pour l'ensemble de la base de données est de 85 pixels (la largeur allant de 80 à 90 pixels, selon la forme et l'étirement de la bouche). Pour chacune des séquences, des images de bouches ouvertes et fermées sont visibles et la bouche est ouverte pour 94 des 150 images de la base**.**

Fig. 4.21. Système d'acquisition de la base Gipsa.

Fig. 4.22. Les six locuteurs de la base Gipsa.

4.4.2. Base AR

La base de visages AR a été créée par A. Martinez et R. Benavente [Martinez, 1998] au Computer Vision Center de l'université U.A.B de Barcelone. La base complète contient plus de 3200 images

statiques correspondant à 126 sujets différents (70 hommes et 56 femmes). Les visages sont vus de face avec différentes expressions faciales, différentes conditions d'illumination et des occultations. Aucune restriction sur la tenue (vêtements, lunette...), sur le maquillage ou sur la coupe de cheveux n'est imposée. Chaque sujet a participé à deux sessions d'acquisition, espacées de 14 jours.
La base AR est disponible à l'adresse suivante : http://cobweb.ecn.purdue.edu/~aleix/ar.html. Les images sont en RGB et de taille 768x576 pixels. La largeur moyenne des bouches pour l'ensemble de la base de données est de 110 pixels (la largeur allant de 100 à 120 pixels, selon la forme et l'étirement de la bouche).
Pour chaque participant, nous avons 26 images correspondant à 13 caractéristiques et 2 sessions différentes. Les caractéristiques sont les suivantes :

1 = expression neutre (BF) 2 = sourire (**BO**)
3 = colère (BF) 4 = cri (**BO**)
5 = lumière venant de la gauche (BF) 6 = lumière venant de la droite (BF)
7 = éclairage uniforme (BF) 8 = port de lunette de soleil (BF)
9 = lunette de soleil + lumière venant de la gauche (BF)
10 = lunette de soleil + lumière venant de la droite (BF)
11 = port d'une écharpe (Bouche non visible)
12 = port d'un écharpe + lumière venant de la gauche (Bouche non visible)
13 = port d'un écharpe + lumière venant de la droite (Bouche non visible)
14 à 26 = session numéro 2 (mêmes conditions que de 1 à 13).

BF = Bouche Fermée
BO = Bouche Ouverte

Dans cette thèse, nous nous intéressons aux images où la bouche est visible. La bouche est ouverte pour les caractéristiques 2, 4, 15, 17 (« sourire » et « cri » pour deux sessions) et ainsi, la base contient 126x2x2 = 504 images de bouches ouvertes. En réalité, nous aurons 507 images car 3 images supplémentaires sont disponibles pour la caractéristique 4. La bouche est fermée pour le reste de images où la bouche est visible (images sans écharpe). Pour tester le cas bouche fermée, nous utiliserons les images avec les caractéristiques 1, 3, 8 (session 1) et 14, 16 et 21 (session2), ce qui correspond à 126x3x2 = 756 images.
La figure 4.23 montre des exemples d'images de la base AR où la bouche est fermée pour les deux sessions.

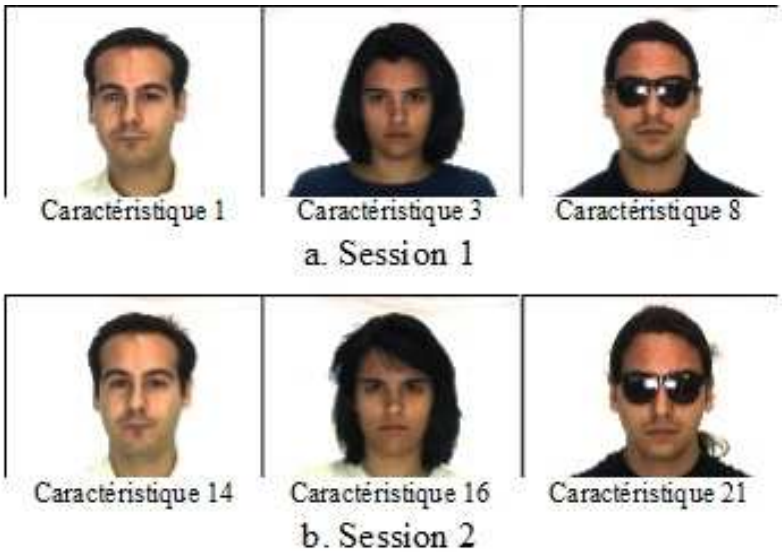


Fig. 4.23. Exemples d'images de la base AR [Martinez, 1998] où la bouche est fermée.

Les figure 4.24 et 4.25 montrent des exemples d'images pour les caractéristiques « sourire » et « cri ».

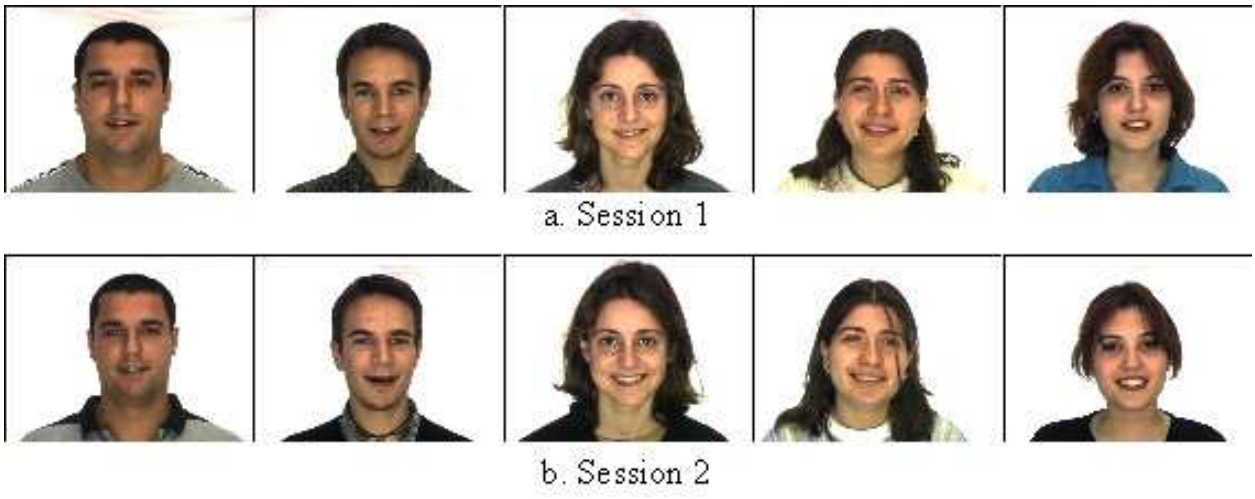


Fig. 4.24. Exemples d'images de la base AR [Martinez, 1998] pour la caractéristique « sourire ».

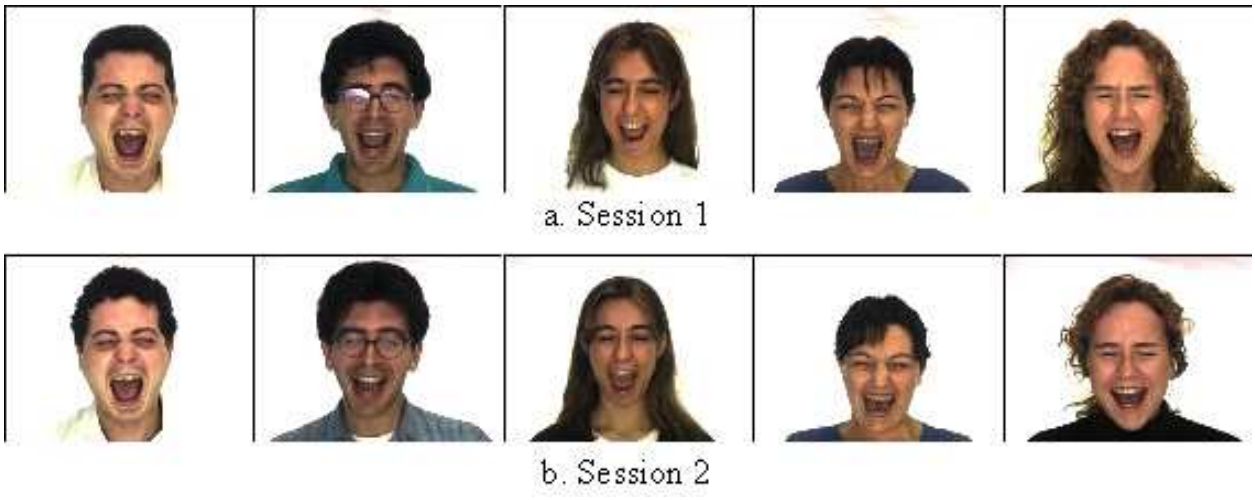


Fig. 4.25. Exemples d'images de la base AR [Martinez, 1998] pour la caractéristique « cri ».

4.4.3. Base TELMA

Dans le cadre du projet TELMA (cf. Section 1.2), une campagne d'acquisition de séquences vidéo a été réalisée en 2005 dans le contexte du langage LPC (cf. Section 1.2.1). La base TELMA contient 250 séquences de longueurs variables (allant de 79 à 1412 images) représentant 82530 images. Les séquences montrent une femme vue de face et codant des phrases en langage LPC. A chaque séquence correspond une phrase différente.

L'enregistrement vidéo a été effectué au laboratoire Gipsa, au sein du Département Parole et Cognition, à une cadence de 50 images/seconde. Les images sont au format BMP et de taille 720x576 pixels. Elles ont été acquises en plan large, afin de filmer à la fois le visage et la main du locuteur. La largeur moyenne des bouches pour l'ensemble de la base de données est de 50 pixels. Les conditions d'éclairage et d'acquisition (caméra et distance du visage par rapport à la caméra) sont les mêmes pour les 250 séquences. La codeuse ne porte aucun maquillage particulier et elle est libre de bouger sa tête.

La figure 4.26 montre des exemples d'images de la base TELMA sur lesquelles il est clair que la taille de la bouche est réduite.

Fig. 4.26. Exemples d'images de la base TELMA.

CHAPITRE 5

Segmentation statique des lèvres

L'algorithme de segmentation, que nous proposons dans ce chapitre, permet d'extraire les contours intérieur et extérieur des lèvres dans des images statiques, nous le nommerons donc **algorithme statique**. Le schéma global du fonctionnement de l'algorithme statique est présenté sur la figure 5.01, il montre l'enchaînement des 3 phases de la méthode (1S, 2S, 3S).

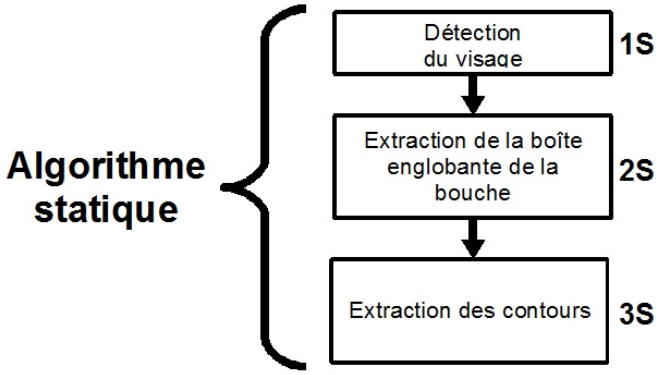


Fig. 5.01. Schéma global de la segmentation statique des contours des lèvres.

Dans ce chapitre, nous présentons les 3 phases de la segmentation statique de la manière suivante. La détection du visage (1S) et le calcul de la boîte autour de la bouche (2S) sont détaillés dans la partie 5.1. Cette étape permet de générer une région d'intérêt et de réduire les coûts de calcul.

La partie 5.2 décrit le processus de segmentation du contour extérieur. La méthode d'extraction est une amélioration de l'algorithme développé par Eveno dans [Eveno, 2004]. D'une part l'obtention du contour extérieur est réalisée de manière automatique (dans [Eveno, 2004], l'initialisation était manuelle), d'autre part, la détection de points clefs sur le contour extérieur est rendue plus robuste grâce à la convergence de deux jumping snakes (cf. Chapitre 4) au lieu d'un seul pour la méthode originale. Le modèle paramétrique extérieur composé de 4 cubiques est positionné à partir de ces points clefs et l'optimisation de ce modèle, permettant l'extraction des contours extérieurs de la bouche, est réalisé en maximisant les flux moyens des gradients introduits dans la section 4.3 pour accentuer le contour des lèvres.

Dans la partie 5.3, le même type de traitement est adapté au contour intérieur des lèvres. Une étape supplémentaire donnant l'information sur l'état de la bouche (ouverte ou fermée) est nécessaire pour choisir le modèle paramétrique intérieur adéquat. Par ailleurs, en raison de la grande variabilité de l'apparence de l'intérieur de la bouche, les points clefs internes doivent être ajustés.

Finalement, dans la partie 5.4, nous effectuons une évaluation quantitative de l'algorithme de segmentation proposé.

5.1. Initialisation de l'algorithme statique

La première étape d'un algorithme de segmentation est la recherche d'une zone d'intérêt pour réduire la zone de recherche des contours. De plus, cela permet de diminuer les coûts de calcul liés au traitement d'image (calcul des espaces couleurs, des gradients...). Dans le contexte de la détection des contours des lèvres, il faut tout d'abord localiser le visage et ensuite se focaliser sur la bouche.

5.1.1. Détection du visage : algorithme CFF (Phase 1S)

De nombreuses études ont permis de proposer des méthodes robustes de détection du visage; une comparaison de plusieurs algorithmes de détection est réalisée dans [Yang, 2002]. Nous avons décidé d'utiliser l'algorithme CFF (**C**onvolutional **F**ace **F**inder), développé par Christophe Garcia [Garcia, 2004] à France Telecom R&D. Dans le cadre du projet TELMA, cet algorithme a été utilisé dans le travail de thèse de Thomas Burger [Burger, 2007]. En outre, le CFF propose un taux de détection performant sur les bases de visage standards (cf. Fig. 5.02).

Face Detector	CMU	CMU-125	MIT	MIT-20
Colmenarez and Huang	93.9%/8122			
Féraud *et al.*	86.0%/8			
Yang *et al.*		93.6%/74		91.5%/1
Osuna *et al.*			74.2%/20	
Roth *et al.*		94.8%/78		94.1%/3
Rowley *et al.*	86.2%/23		84.5%/8	
Schneiderman and Kanade		94.4%/65		
Sung and Poggio			79.9%/5	
Viola and Jones	88.4%/31		77.8%/5	
Li *et al.*	90.2%/31			
Convolutional Face Finder	90.3%/8	90.5%/8	90.1%/7	90.2%/5

Fig. 5.02. Performances de l'algorithme CFF [Garcia, 2004].

L'algorithme CFF est une approche basée sur des réseaux de neurones convolutionnels (**C**onvolutional **N**eural **N**etwork, CNN). Le principe de fonctionnement du CFF est le suivant : une image en niveau de gris est proposée en entrée du CNN et celui-ci calcule une série de descripteurs de l'image à partir d'une succession de convolutions, puis effectue une classification en fonction de ces descripteurs. En sortie, l'algorithme fournit les coordonnées de la boîte englobante du visage (cf. Fig. 5.03).

Le CFF est un algorithme robuste qui permet la détection de plusieurs visages dans une même image avec des tailles et des apparences variables. De plus, il autorise certaines libertés de mouvement pour les visages. Ainsi, les performances de l'algorithme de détection CFF ne sont pas altérées si la tête effectue des rotations allant de - 20 à + 20 degrés par rapport au plan de l'image (Angle *pan* qui correspond au mouvement de la tête qui se tourne de la gauche vers la droite) ou des inclinaisons de côté allant de - 60 à + 60 degrés (Angle *roll* qui correspond au mouvement de la tête qui se penche vers la gauche ou vers la droite).

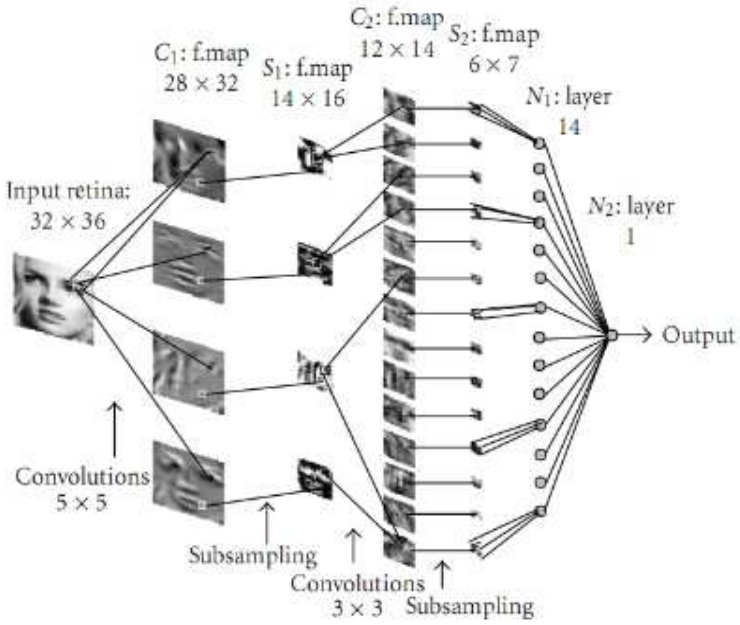


Fig. 5.03. Architecture de l'algorithme CFF [Garcia, 2004].

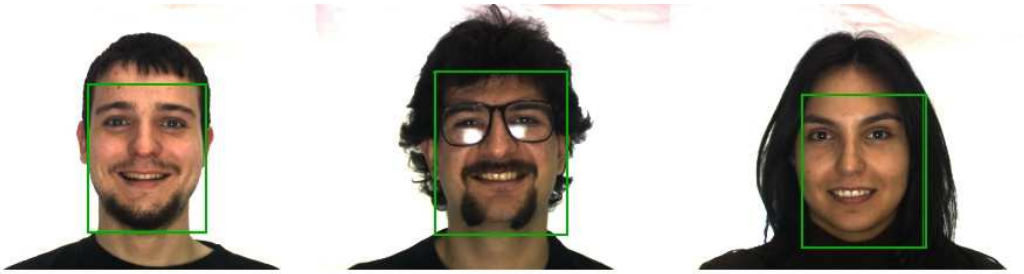
En ce qui nous concerne, ces contraintes de rotations et d'inclinaisons sont respectées dans le contexte applicatif étudié et dans les différentes bases d'images que nous utilisons pour développer et tester notre algorithme de segmentation des lèvres.

- Le logiciel Makeuponline nécessite que la personne, qui désire se maquiller virtuellement, se présente de face (cf. Section 1.1.2).
- Dans les images de la base TELMA (cf. Section 4.4), la personne qui code en LPC est face à la caméra.
- La base AR [Martinez, 1998] est constituée uniquement d'images où le visage est vu de face (cf. Section 4.4).
- Enfin en ce qui concerne les images du laboratoire (base Gipsa, cf. Section 4.4), elles sont déjà centrées sur la bouche (dans ce cas, la détection du visage n'est pas possible et la boîte autour de la bouche est choisie manuellement).

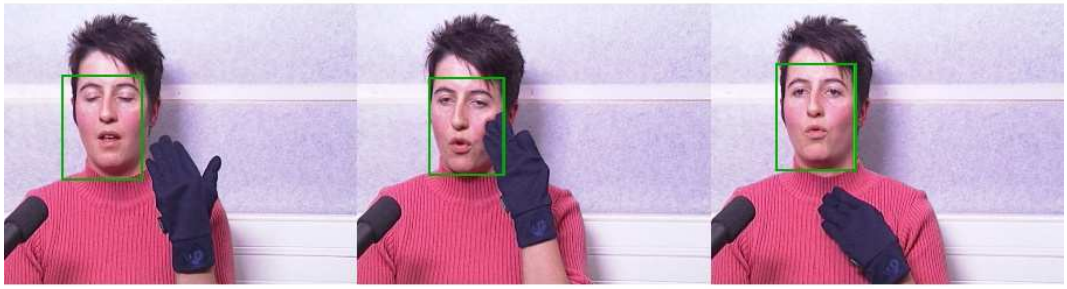
Aussi, les résultats obtenus par le CFF sont d'autant plus performants que sur chaque image que nous segmentons, il n'y a qu'un seul visage, qui est l'objet principal de l'image.
La figure 5.04 montre un éventail des images que nous avons à disposition avec le cadre obtenu autour du visage.

Comme nous pouvons le voir sur la figure 5.04, la boîte autour du visage, obtenue avec l'algorithme CFF, n'est pas toujours centrée sur le visage. Cependant les yeux, le nez et surtout la bouche sont toujours à l'intérieur de celle-ci. Ceci va nous permettre de détecter la région de la bouche.

A partir des résultats du CFF, nous créons une nouvelle image à partir des limites de la boîte encadrant le visage (cf. Fig. 5.05). Cela permet de réduire les temps de calcul des différents espaces couleurs nécessaires à nos gradients définis au chapitre 4.

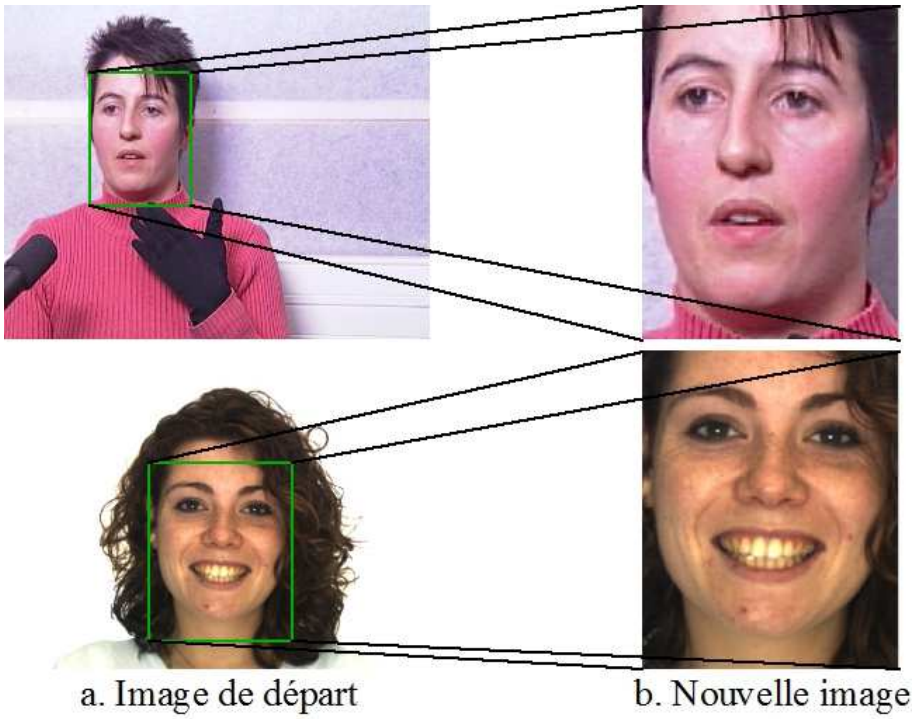


Images de la base AR [Martinez, 1998]



Images du projet TELMA

Fig. 5.04. Résultats CFF sur des exemples d'images testées pour la segmentation.



a. Image de départ b. Nouvelle image

Fig. 5.05. Nouvelle image à partir des limites de la boîte englobante du visage.

5.1.2. Détection de la boîte englobante de la bouche (Phase 2S)

La détection de la boîte englobante de la bouche est plus difficile. Cependant, la boîte sert, pour la suite de l'algorithme, à limiter la recherche des contours des lèvres à la zone définie par la boîte englobante de la bouche et à définir la position des 2 germes pour le jumping snake extérieur supérieur et le jumping snake extérieur inférieur. Or, ces germes ne sont soumis qu'à peu de contraintes pour leur

positionnement; ils doivent être situés plus proche de la bouche que du nez ou du menton (cf. section 4.2). Ainsi, la boîte englobante de la bouche n'a pas besoin d'être très précise, mais elle doit contenir entièrement la bouche.

Détection d'une 1ère boîte englobante :

Dans [Duffner, 2005], Garcia *et al.* ont proposé une extension de l'algorithme CFF pour extraire des points caractéristiques du visage appelé **C**onvolutional **F**ace and **F**eatures **F**inder (C3F). Le C3F fonctionne sur le même principe que le CFF et donne en sortie la position des centres des yeux, du centre du nez et du centre de la bouche (cf. Fig. 5.06.a).

Une première boîte englobante de la bouche est calculée en définissant les 4 côtés directement à partir des 4 points donnés par le C3F. Soient (x_{le}, y_{le}), (x_{re}, y_{re}), (x_n, y_n) et (x_m, y_m) les coordonnées du centre de l'œil gauche, du centre de l'œil droit, du centre du nez et du centre de la bouche.
Soit $d_{le\text{-}re}$, la distance entre les 2 centres des yeux, les coordonnées de la boîte autour de la bouche sont définies par :

- Les ordonnées du côté haut sont égales à $y_n+[(y_m- y_n)/2]$ (pour être situé entre le centre du nez et le centre de la bouche)
- Les ordonnées du côté bas sont égales à $y_m+d_{le\text{-}re}$
- Les abscisses du côté gauche sont égales à $x_m-(0{,}6*d_{le\text{-}re})$
- Les abscisses du côté gauche sont égales à $x_{m+}(0{,}6*d_{le\text{-}re})$

Le résultat de cette 1ère boîte permet d'avoir un cadre contenant obligatoirement la bouche. Le côté bas n'est en général pas proche de la bouche, mais celui-ci est choisi intentionnellement beaucoup plus bas pour être en dessous de la bouche dans tous les cas possibles : fermée, ouverte et largement ouverte (cf. Fig. 5.06.b).

Ajustement de la 1ère boîte :

Pour avoir un côté bas plus proche de la bouche, il faut réaliser un ajustement. Nous l'effectuons en faisant une statistique des couleurs du contenu de la 1ère boîte et en procédant à un seuillage. Pour cela, nous utilisons la somme normalisée entre 0 et 1 des pseudo-teintes normalisées $H1_N + H2_N$ (cf. Section 4.3), ce qui accentue le contraste entre les lèvres et la peau, car dans la 1ère boîte, il n'y a en principe que la bouche et de la peau. Nous avons donc une imagette (sous image de $H1_N + H2_N$ définie par le cadre de la 1ère boîte englobante, cf. Fig. 5.06.c) en niveau de gris. Un seuillage est réalisé en utilisant la méthode d'Otsu [Otsu, 1979] qui permet de convertir une image en niveau de gris en une image binaire par choix d'un seuil qui minimise la variance intraclasse (cf. Fig. 5.06.d).
Finalement, le côté bas est remonté pour se rapprocher de la région seuillée (cf. Fig. 5.06.e).

Il est possible que le seuillage ne soit pas efficace et que la région obtenue soit trop grande, c'est le cas du dernier exemple de la figure 5.06. Si la région obtenue après le seuillage est constituée d'un nombre de pixels trop grand (nous avons fixé la condition à < ½ * le nombre de pixels situés à l'intérieur de la 1ère boîte), le côté bas est ajusté à la position $y_m+[(y_m- y_n)/2]$ (cf. 2 derniers cas de la figure 5.06.e). Cette valeur tient compte de la façon dont sont trouvées les ordonnées du côté haut et fonctionne pour toutes les images que nous avons testées.

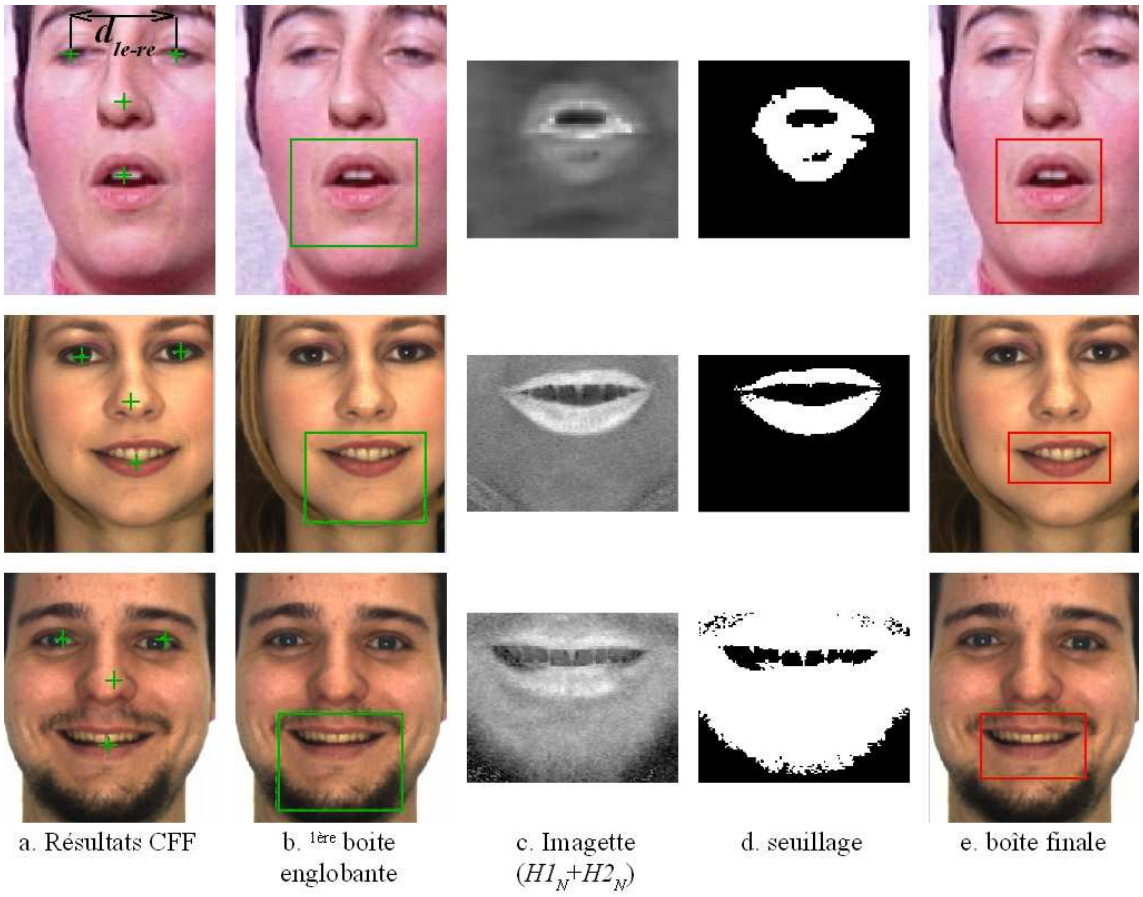


Fig. 5.06. Principe de la détection de la boîte englobante de la bouche.

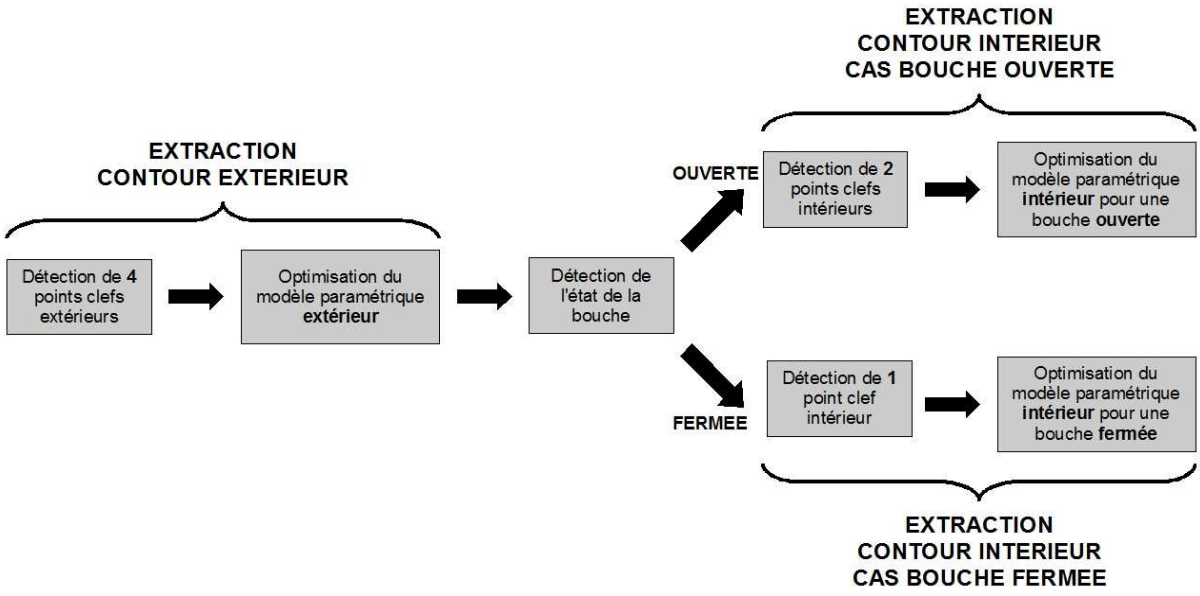


Fig. 5.07. Déroulement de l'extraction des contours (algorithme statique).

Maintenant que la boîte englobante de la bouche est trouvée, nous pouvons commencer à extraire les contours des lèvres dans l'image (Phase 3S). Le schéma de la figure 5.07 montre le déroulement de la segmentation. La section 5.2 détaille les différentes étapes nécessaires à l'extraction du contour extérieur de la bouche pour une image statique (détection de plusieurs points clefs externes, positionnement et optimisation du modèle extérieur des lèvres). Ensuite, nous pouvons passer à la segmentation de l'intérieur de la bouche, dont le processus est décrit dans la partie 5.3 (information sur l'état de la bouche, détection et ajustement de 2 points clefs internes, positionnement et optimisation du modèle intérieur des lèvres).
Pour la segmentation du contour extérieur et la segmentation du contour intérieur, nous utilisons la même stratégie qui consiste à combiner les jumping snakes et les modèles paramétriques.
Dans les deux cas, les jumping snakes sont utilisés dans l'étape d'initialisation (détection de la position des points clefs externes et internes), alors que les modèles paramétriques ont été construits pour modéliser le contour labial (l'optimisation des modèles permet d'extraire les contours des lèvres).

5.2. Extraction du contour extérieur

Dans cette partie, nous allons voir comment l'algorithme jumping snake peut être utilisé pour placer 4 points clefs extérieurs (P_2, P_3, P_4 et P_6) qui permettent d'initialiser la position du modèle paramétrique extérieur. Ensuite, la détection des commissures (P_1 et P_5) et le calcul des courbes cubiques du modèles (γ_1, γ_2, γ_3 et γ_4) se fait en une seule étape.
Notre méthode d'extraction du contour extérieur est une amélioration de l'algorithme développé par Eveno dans [Eveno, 2004]. En effet, d'une part l'obtention du contour extérieur est réalisée de manière automatique, alors que dans [Eveno, 2004], un germe doit être placé manuellement au dessus de la bouche. D'autre part, nous verrons que nous avons ajouté un deuxième snake, pour une détection plus robuste du point P_6, et que nous avons utilisé les gradients introduits dans la section 4.3.

5.2.1. Détection des points clefs extérieurs

Le modèle paramétrique extérieur présenté dans la partie 4.1 est composé de 6 points clefs $P_{i=1 \text{ à } 6}$ (cf. Fig. 5.08). Les points P_2, P_3, P_4 et P_6 peuvent être localisés directement en faisant converger 2 jumping snakes. Les commissures extérieures des lèvres (P_1 et P_5) sont plus difficiles à détecter localement; elles seront déterminées en même temps que le calcul des courbes du modèle (cf. Partie 5.2.2).

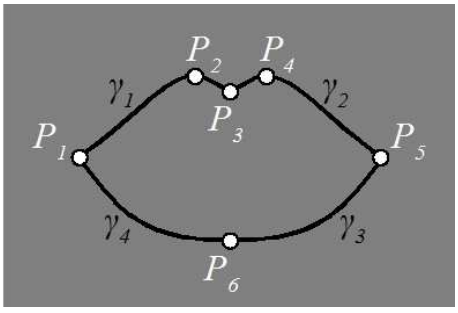


Fig. 5.08. Rappel : modèle paramétrique extérieur.

Un jumping snake extérieur supérieur est utilisé pour les points extérieurs hauts P_2, P_3 et P_4, alors qu'un jumping snake extérieur inférieur sert à trouver P_6. Nous avons vu que l'algorithme jumping snake est initialisé par un seul point (le germe). La boîte englobante de la bouche trouvée précédemment permet d'initialiser ces 2 snakes. Les 2 germes sont choisis comme étant les points milieux des côtés haut et bas de la boîte englobante (cf. Fig. 5.08). En effet, ces positions respectent les contraintes sur la position des germes présentées dans la partie 4.2.3.a.

Pour rappel, il faut que :

- pour le contour extérieur haut, le germe se trouve au dessus de la bouche et il doit être plus proche de la bouche que du nez;
- pour le contour extérieur bas, le germe doit se trouver en dessous de la bouche et il doit être plus proche de la bouche que du menton.

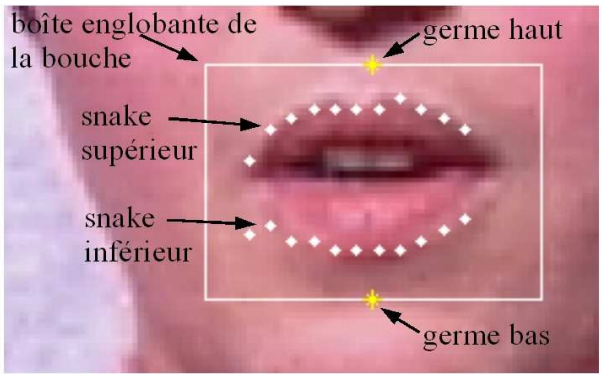


Fig. 5.09. Extraction du contour extérieur : jumping snakes.

- Pour le snake supérieur : les paramètres sont réglés de façon à ce que le snake se propage en dessous du germe supérieur (N_{up}=3 < N_{low}=10, cf. partie 4.2) et le gradient G_1 (cf. partie 4.3) est utilisé pour sa convergence.
- Pour le snake inférieur : les paramètres sont réglés de façon à ce que le snake se propage en dessus du germe inférieur (N_{up}=10 > N_{low}=3) et le gradient G_2 est utilisé pour sa convergence.

Une fois que les 2 snakes ont convergé, nous obtenons des points sur les contours extérieurs supérieur et inférieur (cf. Fig. 5.09). Le snake supérieur donne la position des 3 points de l'arc Cupidon. P_2 et P_4 sont les points les plus hauts du snake respectivement à gauche et à droite de la verticale passant par les 2 germes. P_3 est le point du snake le plus bas situé entre P_2 et P_4. Alors que le point P_6 est le point du snake inférieur le plus proche de la verticale passant par P_3 (cf. Fig. 5.10).

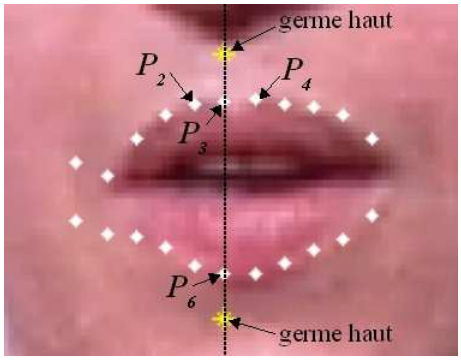


Fig. 5.10. Détection des points clefs extérieurs à partir des points des snakes.

Dans l'algorithme original [Eveno, 2004], le point P_6 est trouvé en analysant la composante verticale du gradient de la pseudo-teinte *H2* le long de la verticale passant par P_3. La position de P_6 est déterminée en deux étapes (cf. [Eveno, 2003] pour plus de détails) :

- un point intermédiaire est associé au maximum du gradient de *H2* (ce point se trouve sur le contour intérieur de la lèvre inférieure (cf. Fig. 5.11.a)),
- le point P_6 est associé au minimum du gradient de *H2* en-dessous du point intermédiaire.

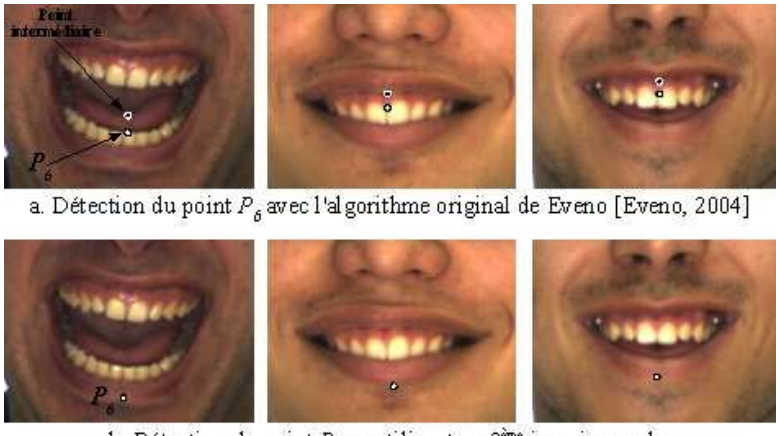
Cette méthode donne des résultats satisfaisants lorsque la bouche est fermée ou légèrement ouverte. Toutefois, des erreurs peuvent se produire à cause de l'apparence sur ou en-dessous de la lèvre inférieure. Les deux premières images de la figure 5.11.a montrent des exemples de mauvaises détections lorsqu'une région sombre est présente au-dessous de la lèvre inférieure; le point P_6 est positionné sur la frontière entre la peau et la zone sombre. Les deux dernières images de la figure 5.11.a montrent des exemples de mauvaises détections lorsqu'une surbrillance est présente sur la lèvre inférieure; le point P_6 est positionné sur la zone surexposée. Ces erreurs s'expliquent par le fait que ces variations d'apparence engendrent des variations importantes sur les valeurs de la pseudo-teinte *H2*, et le minimum du gradient de *H2* ne se trouve plus au niveau du contour labial. La figure 5.11.b montre que la détection est améliorée lorsqu'on utilise un deuxième jumping snake qui converge en utilisant le gradient G_2 (en supposant que la position du germe respecte les contraintes d'initialisation).

a. Détection du point P_6 avec l'algorithme original de Eveno [Eveno, 2004]

b. Détection du point P_6 en utilisant un 2ème jumping snake

Fig. 5.11. Amélioration de la détection du point P_6. Images issues de [Martinez, 1998].

En plus des erreurs de détection dues à l'apparence, la détection du point P_6 avec la méthode originale est difficile lorsque la bouche est ouverte et que la langue ou les gencives sont visibles. Dans ce cas, le point intermédiaire peut se retrouver relativement loin de la lèvre inférieure sur la langue ou sur la frontière entre la lèvre supérieure et les gencives (cf. Fig. 5.12.a). De ce fait, le point P_6 est mal positionné et il se retrouve à l'intérieur de la bouche au niveau des dents (les valeurs de la pseudo-teinte *H2* étant très faibles pour les pixels dent (cf. Section 4.3)). L'utilisation d'un deuxième jumping snake pour le point bas permet de rendre la détection plus robuste (cf. Fig. 5.12.b).



a. Détection du point P_6 avec l'algorithme original de Eveno [Eveno, 2004]

b. Détection du point P_6 en utilisant un 2ème jumping snake

Fig. 5.12. Amélioration de la détection du point P_6 pour des bouches ouvertes. Images issues de [Martinez, 1998].

5.2.2. Détection des commissures et optimisation du modèle paramétrique extérieur

Nous avons donc maintenant la position de 4 points du modèle extérieur et les 2 segments [P_2P_3] et [P_3P_4]. Il nous reste à trouver les commissures (P_1 et P_5) et à calculer les 4 cubiques γ_1, γ_2, γ_3 et γ_4.

La détection des commissures est un challenge difficile dans la mesure où elles ne sont pas vues comme 2 points mais plutôt comme une zone sombre située à chacune des extrémités de la bouche. En effet, sur la figure 5.13.b, l'image est focalisée sur la commissure droite de la bouche et il est difficile de la détecter d'un point de vue local. Lorsque cette commissure est vue de plus loin, il suffit de prolonger les frontières hautes et basses des lèvres pour trouver la commissure à l'intersection de ces 2 contours (illustration sur la figure 5.13.c).

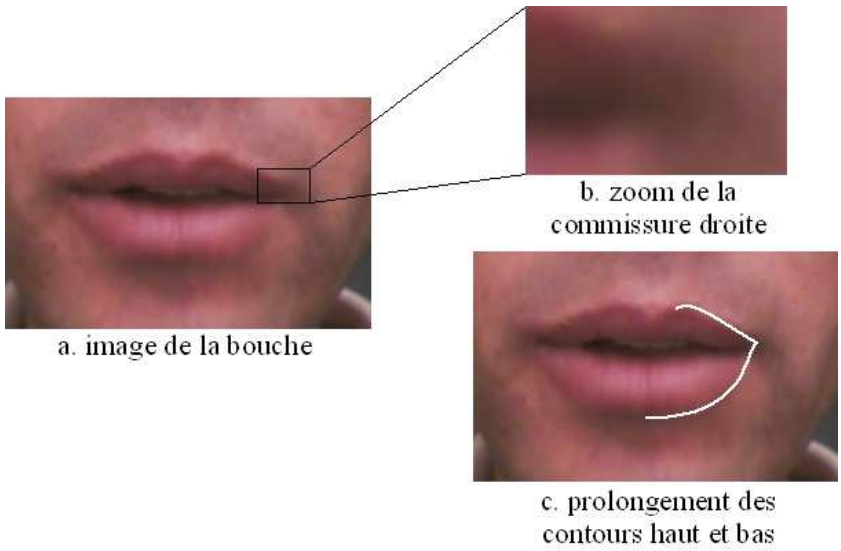


Fig. 5.13. Difficulté de la détection des commissures.

En conséquence, le positionnement des commissures peut être directement lié aux allures des courbes représentant les contours des lèvres. Nous allons donc déterminer P_1 et P_5, et calculer les 4 cubiques du modèle extérieur, en une seule et même opération.

L'expression analytique d'une courbe cubique est donnée par l'équation 5.01.

$$y = ax^3 + bx^2 + cx + d \tag{5.01}$$

Il suffit de 4 équations pour déterminer complètement la courbe polynomiale de degré 3. Dans la description du modèle paramétrique extérieur de la section 4.1, nous avons vu que les 4 cubiques γ_1, γ_2, γ_3 et γ_4 relient les 6 points clefs externes $P_{i=1\ à\ 6}$ (cf. Fig. 4.01) et que nous imposons, pour notre modèle, que les dérivées des cubiques s'annulent en P_2, P_4 et P_6. En conséquence, pour chaque cubique, nous avons un des 2 points d'extrémité (les positions de P_2, P_4 et P_6 ont été trouvées dans la partie 5.2.1) et une contrainte d'annulation de dérivée, ce qui nous donnent 2 équations. En théorie, il suffit donc de connaître encore 2 points appartenant à chaque cubique pour trouver la courbe.

Les 2 jumping snakes extérieurs supérieur et inférieur fournissent des points supplémentaires (cf. Fig. 5.10). Les points des snakes sont fiables au milieu de la bouche, là où les contours sont les plus marqués et où les gradients sont donc plus forts. Alors qu'en se rapprochant des extrémités de la bouche, les points des snakes peuvent ne plus être exactement sur les contours si ceux-ci ne sont pas assez marqués. Ainsi, nous ne pouvons utiliser que les points proches des points clefs P_2, P_4 et P_6 et donc loin des commissures.

En utilisant 2 points supplémentaires, les courbes cubiques peuvent être directement calculées avec maintenant 4 équations ou nous pouvons utiliser la méthode des moindres carrés en utilisant 3 points supplémentaires. Ces 2 méthodes fournissent des résultats approximatifs pour les cubiques et la détection des commissures (aux intersections des cubiques), car on utilise des points trop éloignés des commissures. Aussi, en faisant varier, même légèrement, la position d'un des points, ces méthodes fournissent des résultats très différents.
Pour améliorer la détection, il faudrait obtenir un point plus proche des commissures pour chacune des 4 cubiques. Pour cela, nous allons tout simplement supposer que les 2 commissures (P_1 et P_5) sont connues. En utilisant la position des points clefs P_2, P_4 et P_6, les contraintes sur les dérivées, les points supplémentaires fournis par les snakes et la position des commissures, les cubiques sont calculées rapidement; la méthode des moindres carrés devenant une simple régression linéaire.

Le processus d'optimisation du modèle extérieur et de la détection des commissures est le suivant :

1) Plusieurs pixels candidats sont testés pour trouver les points P_1 et P_5
2) Pour chacun des pixels candidats P_1 (resp. P_5), le couple de cubiques (γ_1 et γ_4) à gauche de la bouche (resp. le couple de cubique (γ_2 et γ_3) à droite de la bouche) est calculé en utilisant les informations citées précédemment.
3) Un critère de maximisation du flux moyen de gradient permet de déterminer le meilleur couple pour chaque côté de la bouche et de trouver P_1 et P_5.

La détermination des positions des commissures et l'optimisation du modèle se font donc en une seule et même opération.

Pour limiter le coût de calcul de cette méthode, il faut pouvoir tester uniquement quelques pixels candidats; une recherche dans toute la boîte englobante de la bouche serait trop longue. Une supposition, que l'on retrouve dans plusieurs études sur la segmentation du contour extérieur des lèvres ([Delmas, 2002]; [Eveno 2004]), est que les commissures se trouvent dans des zones sombres proches des lèvres. Il suffit donc de ne tester que les pixels les plus sombres de la boîte englobante. Nous construisons ce que nous appelons la ligne des minima de luminance, notée L_{min}, qui un chaînage des pixels les plus sombres passant par la bouche. La figure 5.14 montre le schéma de construction de L_{min}. A partir d'un point initial, des points sont ajoutés à gauche et à droite en ne testant que les 3 pixels les plus proches et en choisissant le pixel ayant la luminance la plus faible. Pour que L_{min} passe par la bouche et les 2 commissures, le point initial est choisi comme étant le pixel le plus sombre du segment [P_3P_6].

Des exemples de résultats de calcul de L_{min} sont visibles sur la figure 5.15 (pour des images des bases présentées dans la section 4.4). Dans chacune des images, la ligne des minima de luminance passe effectivement par les 2 commissures.

En se limitant à cette ligne et en utilisant les bornes de la boîte, il est possible de ne tester plus que quelques dizaines de pixels (suivant la précision du résultat de la boîte englobante de la bouche).

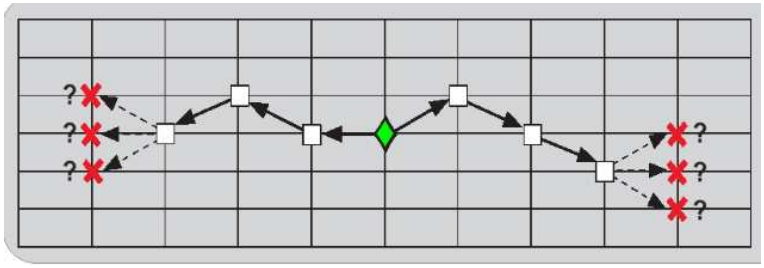
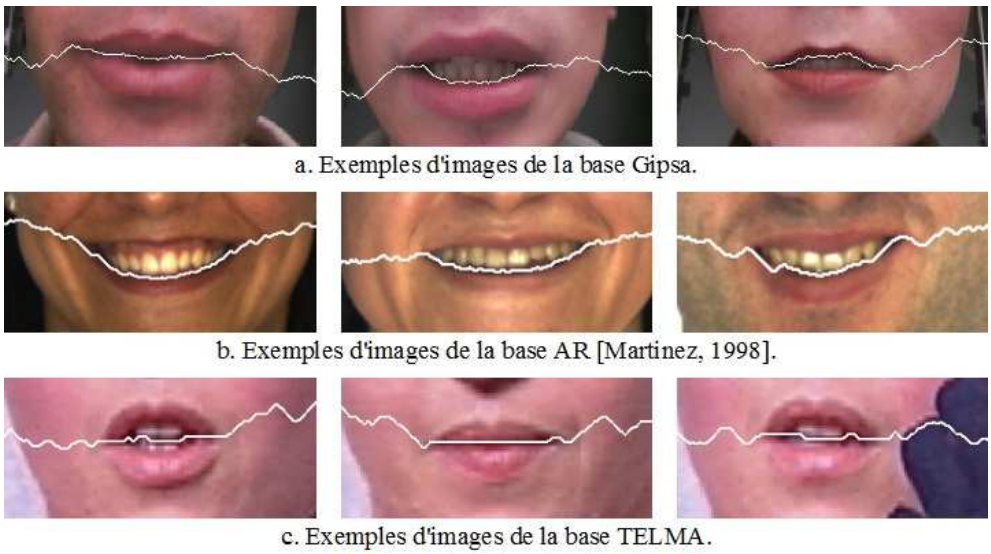


Fig. 5.14. Construction de L_{min}.



a. Exemples d'images de la base Gipsa.

b. Exemples d'images de la base AR [Martinez, 1998].

c. Exemples d'images de la base TELMA.

Fig. 5.15. Exemples de résultats de L_{min} pour des images de bouche.

Maintenant que nous avons les pixels candidats, nous calculons les couples de cubiques ((γ_1, γ_4) et (γ_2 et γ_3)) e t déterminons les meilleurs en utilisant notre critère contour. A la différence de la méthode originale [Eveno, 2003] où les gradients R_{top} et R_{bottom} de l'équation 2.03 sont utilisées pour l'optimisation des modèles, nous utilisons les gradients introduits dans la section 4.3 :

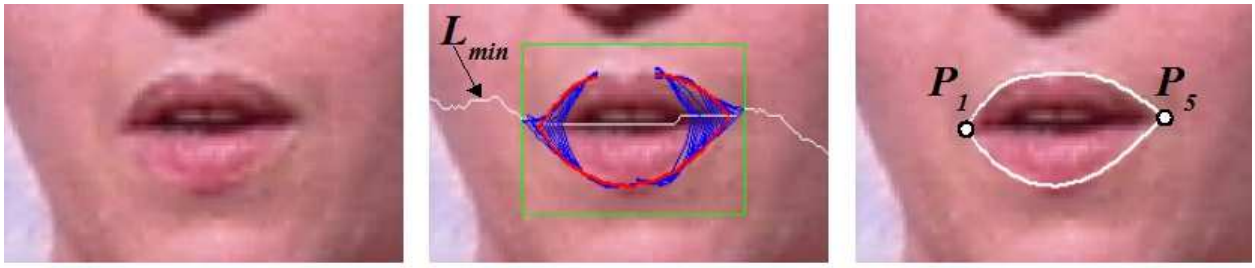
- Pour la cubique γ_1, le gradient G_1 est utilisé en prenant la composante diagonale adéquate (cf. Partie 4.3).
- Pour la cubique γ_2, le gradient G_1 est utilisé en prenant la composante diagonale adéquate
- Pour la cubique γ_3, le gradient G_2 est utilisé en prenant la composante diagonale adéquate
- Pour la cubique γ_4, le gradient G_2 est utilisé en prenant la composante diagonale adéquate

Le flux Φ_i à travers la cubique γ_i est :

$$\Phi_i = \frac{\int_{\gamma_i} G_{1ou2}.dn}{\int_{\gamma_i} ds} \tag{5.02}$$

où *dn* est le vecteur orthogonal au contour et *ds* est l'abscisse curviligne.

Pour chaque pixel candidat P_1 (resp. P_5), la somme $\Phi_1+\Phi_4$ (resp. $\Phi_2+\Phi_3$) est calculée et la somme la plus grande détermine les 2 cubiques du modèle et la position de la commissure. La figure 5.16 décrit le processus d'optimisation du modèle extérieur.



a. Image de la bouche b. Test des pixels candidats aux commissures c. Détection de P_1 et P_5

Fig. 5.16. Optimisation du modèle extérieur : calcul des cubiques et détection des commissures.

Une évaluation de l'algorithme statique pour l'extraction du contour extérieur des lèvres est présentée à la fin de ce chapitre.

5.3. Extraction du contour intérieur

La segmentation du contour intérieur est basée sur le même principe que la segmentation du contour extérieur : des points clefs internes sont détectés suite à la convergence d'un jumping snake pour positionner le modèle paramétrique intérieur et l'optimisation du modèle est réalisée à l'aide d'un critère contour. Il faut cependant déterminer quel modèle intérieur (parmi celui « bouche ouverte » ou celui « bouche fermée) est approprié pour l'image traitée.

La partie 5.3.1 présente la méthode de détection de l'état de la bouche (ouverte ou fermée) que nous avons développée.

La détection du contour intérieur dans le cas d'une bouche ouverte est détaillée dans la section 5.3.2. Les convergences de 2 jumping snakes permettent de trouver les 2 points clefs internes P_8 et P_{10}. Le modèle interne, composé de 4 cubiques, est ajusté en maximisant les flux moyens des gradients G_3 et G_4 (cf. Chapitre 4).

Dans la partie 5.3.3, nous verrons comment, à partir de la ligne des minima de luminance, nous pouvons placer le modèle interne sur le contour intérieur dans le cas d'une bouche fermée.

5.3.1. Détection de l'état de la bouche

Dans le chapitre 2, nous avons présenté l'état de l'art sur la détection de l'état de la bouche à travers différents travaux traitant de la segmentation des lèvres. Nous avons développé et testé des algorithmes de détection basés sur des méthodes qui analysent le gradient intensité [Zhang, 1997], ou qui affectent l'état de la bouche en fonction du résultat de la projection des colonnes du plan teinte ([Pantic, 2001]; [Yin, 2002]). Cependant, aucune de ces méthodes n'a permis d'obtenir des résultats satisfaisants sur nos bases d'images.
Nous avons donc opté pour une stratégie différente qui consiste à obtenir l'information sur l'état de la bouche a posteriori, c'est-à-dire après la segmentation du contour intérieur.

Avant de commencer l'extraction du contour intérieur, nous supposons que la bouche est ouverte et nous procédons donc à la segmentation en utilisant le modèle paramétrique intérieur pour le cas d'une bouche ouverte. A la fin du processus, nous vérifions cette hypothèse et 2 cas se distinguent :

1) La bouche était bien ouverte et la segmentation du contour intérieur est terminée;
2) La bouche était fermée et nous recommençons la segmentation en utilisant le modèle paramétrique intérieur pour le cas d'une bouche fermée.

Cette stratégie est résumée sur la figure 5.17. Le déroulement de l'algorithme statique tel que présenté sur la figure 5.07 est modifié dans la mesure où l'information sur l'état de la bouche est connue après la segmentation du contour intérieur (cas bouche ouverte).

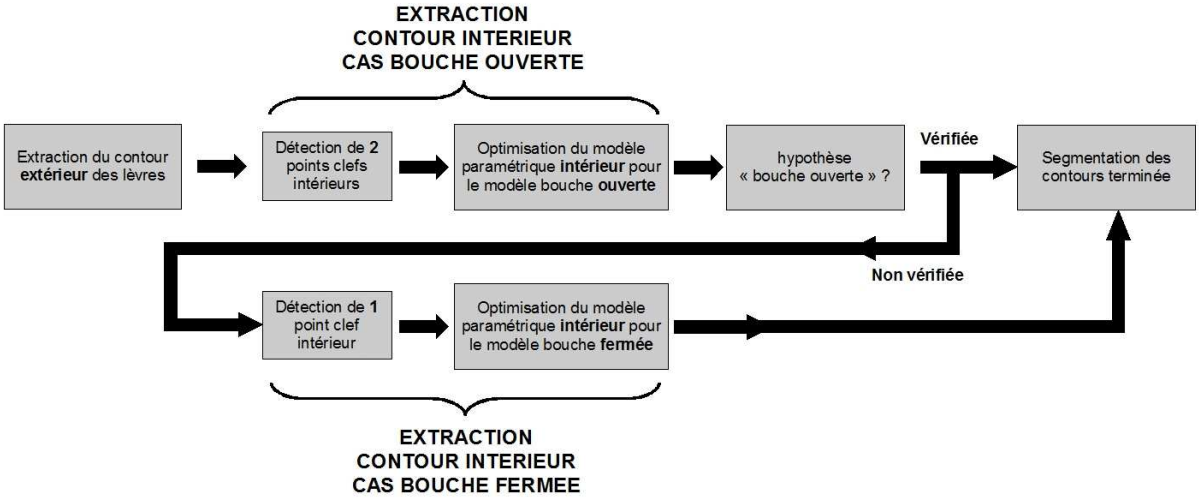


Fig. 5.17. Déroulement de la segmentation du contour intérieur.

La méthode de vérification de l'hypothèse est détaillée dans la partie suivante. D'ores et déjà, signalons que si la bouche est, en réalité, fermée, la région interne obtenue après segmentation, sera petite et plate. Ainsi, nous avons développé une méthode de vérification basée sur un critère géométrique analysant la forme de la région définie par le contour intérieur trouvé (cf. Partie 5.3.2).

5.3.2. Segmentation du contour intérieur : cas bouche ouverte

La méthode d'extraction du contour intérieur dans le cas de la bouche ouverte suit les mêmes étapes que pour le contour extérieur. Deux jumping snakes donnent la position des 2 points clefs internes P_8 et P_{10} (cf. Paragraphe 5.3.2.a), le modèle paramétrique intérieur est initialisé et ajusté à l'aide d'un critère contour (cf. Paragraphe 5.3.2.b). La finalisation de l'extraction est déterminée par la vérification de l'hypothèse « bouche ouverte » (cf. Paragraphe 5.3.2.c). Si la bouche est en fait fermée, le cas bouche est fermée est traitée (cf. Partie 5.3.3).

Pour rappel, le modèle dans le cas d'une bouche ouverte est composé de 4 courbes cubiques et de 4 points clefs (cf. Fig. 5.18).

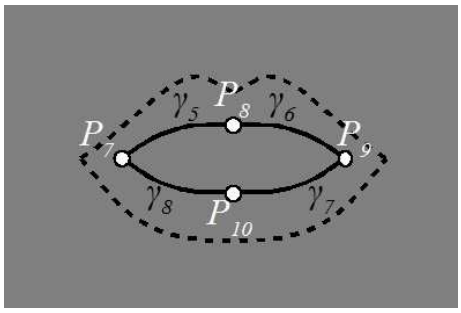


Fig. 5.18. Rappel : Modèle paramétrique intérieur (bouche ouverte).

5.3.2.a. Initialisation des snakes intérieurs et détection des points clefs internes

Positionnement des germes :

La position du point clef P_8 est trouvée en faisant converger un snake intérieur supérieur et celle de P_{10} en faisant converger un snake intérieur inférieur. En effet, P_8 et P_{10} sont respectivement les points milieux des contours internes supérieur et inférieur. Les 2 germes doivent respecter les contraintes établies dans la partie 4.2.3.a :

- Le germe intérieur haut est au dessus du contour intérieur supérieur et plus proche de celui-ci que du contour extérieur supérieur;
- Le germe intérieur bas est au dessous du contour intérieur inférieur et plus proche de celui-ci que du contour extérieur inférieur.

Ainsi les zones d'initialisation des germes sont relativement grandes et nous n'avons pas besoin d'une grande précision de placement, tant que ces contraintes sont satisfaites.

Connaissant la position du point P_3, qui se trouve au milieu de l'arc de Cupidon (et donc au milieu de la bouche lorsque le visage est vu de face), les abscisses des germes intérieurs sont affectées à la valeur de l'abscisse de P_3. En ce qui concerne les ordonnées, nous allons étudier les gradients G_3 et G_4 introduits dans la partie 4.3 qui accentuent respectivement les contours intérieurs haut et bas.
Deux points intermédiaires, notés P'_8 et P'_{10} (car ceux-ci peuvent être vus comme une estimation des positions de P_8 et P_{10}), sont positionnés sur le segment $[P_3P_6]$. L'ordonnée de P'_8 (resp. P'_{10}) est choisie au niveau du maximum du gradient G_3 (resp. G_4) entre les points P_3 et P_6. Afin d'éviter une mauvaise affectation à cause du bruit, une accumulation des gradients est réalisée sur 10 colonnes autour de P_3 et nous choisissons les valeurs cumulées maximales. De plus, seule la composante horizontale des gradients est utilisée, car au milieu de la bouche, les contours sont principalement horizontaux. Ainsi, nous n'utilisons que le filtre F_2 (cf. eq. 4.06) pour déterminer la position de P'_8 et P'_{10}.
La figure 5.19 montre un exemple de l'affectation des points P'_8 et P'_{10} en fonction de l'accumulation verticale des gradients G_3 et G_4.

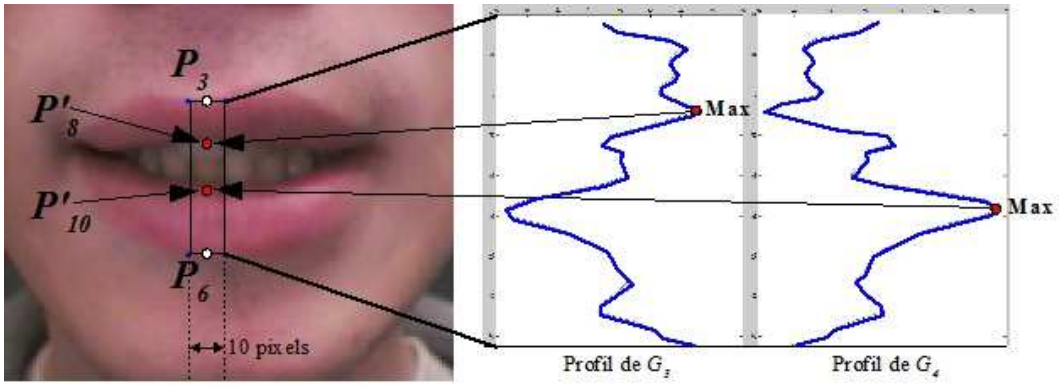


Fig. 5.19. Positionnement des points intermédiaires P'_8 et P'_{10}.

Normalement, ces points sont déjà de bonnes estimations de P_8 et P_{10}, et ils se trouvent sur les contours intérieurs. Toutefois, comme nous l'avons expliqué dans la section 4.3, les gradients G_3 et G_4 ont été construits pour accentuer le contour intérieur dans tous les cas possibles (frontière Lèvre/Dent, Lèvre/Gencive, Lèvre/Langue et Lèvre/Cavité orale). Par conséquent, d'autres contours à l'intérieur de la bouche peuvent être également accentués (on peut le voir sur les figures 4.18 et 4.19), comme par exemple, la frontière Dent/Cavité orale. Il se peut que le point P'_8 ou P'_{10} ne soit pas tout à fait sur le contour des lèvres (cf. Fig. 5.20). Nous verrons plus tard comment ces points peuvent être ajustés en détectant un masque des pixels dents à l'intérieur de la bouche. Les points P'_8 et P'_{10} ne sont donc pas suffisamment fiables pour être choisis directement comme points clefs internes.

Fig. 5.20. Positionnement approximatif des points P'_8 et P'_{10}.

Les 2 germes intérieurs sont positionnés de la manière suivante (cf. Fig. 5.21) :

- Le germe intérieur haut (noté *germe*$_8$ sur la figure 5.21) est placé au ¾ du segment $[P_3P'_8]$,
- Le germe intérieur bas (noté *germe*$_{10}$ sur la figure 5.21) est placé au ¾ du segment $[P_6P'_{10}]$.

De cette façon, les germes respectent les conditions d'initialisation et se trouvent bien sur les lèvres.

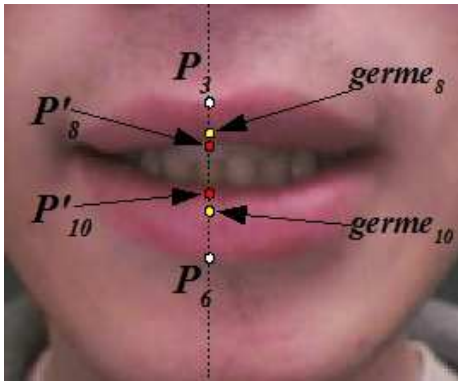


Fig. 5.21. Positionnement des germes intérieurs haut et bas.

Convergence des snakes et détection de P_8 et P_{10} :

Maintenant que nous avons les 2 germes, il est possible de faire converger les snakes intérieurs supérieur et inférieur.

- Pour le snake supérieur : les paramètres sont réglés de façon à ce que le snake se propage en dessous du germe haut (N_{up}=3 < N_{low}=10, cf. partie 4.2) et le gradient G_3 est utilisé pour sa convergence.
- Pour le snake inférieur : les paramètres sont réglés de façon à ce que le snake se propage en dessus du germe bas (N_{up}=10 > N_{low}=3) et le gradient G_4 est utilisé pour sa convergence.

La convergence des snakes donne des points sur les contours intérieurs supérieur et inférieur. P_8 est le point du snake supérieur le plus proche de la verticale passant par P_3 et P_{10} est le point du snake inférieur le plus proche de cette même verticale (cf. Fig. 5.22). Les abscisses des points P_8 et P_{10} sont modifiées pour être égales à l'abscisse de P_3, pour que ces deux points se retrouvent sur la même verticale. Il est à noter que les positions des commissures internes P_7 et P_9 seront trouvées en même temps que l'optimisation du modèle paramétrique intérieur, comme cela a été fait pour les commissures P_1 et P_5.

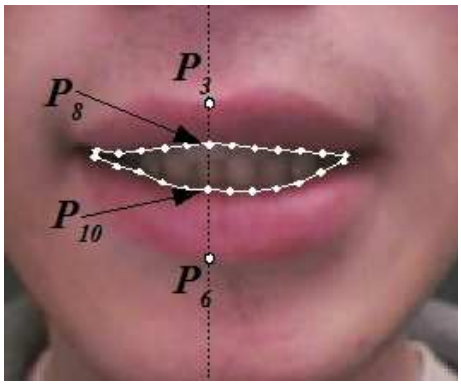


Fig. 5.22. Détection de P_8 et P_{10}.

Ajustement des snakes :

Dans certains cas, il est possible que les snakes ne se stabilisent pas sur les bons contours intérieurs, mais sur d'autres contours parasites, soit parce qu'un autre contour intérieur a été également accentué par les gradients G_3 et G_4, soit parce que le contour n'est pas assez marqué.
Ces mauvaises convergences arrivent essentiellement dans deux cas : à cause de l'accentuation de la frontière Dent/Cavité orale ou en présence des gencives.

Lorsque les dents sont visibles, il arrive que le snake intérieur supérieur ou inférieur se bloque sur le contour situé entre les dents et la cavité orale (cf. Fig. 5.24.b) pour deux raisons :

- ce contour est accentué par le gradient G_{3ou4} et il se trouve trop proche du germe,
- les dents ne sont pas assez brillantes (exemple avec des dents apparaissant plus jaunes que blanches) pour que la composante *u* joue son rôle dans la combinaison des gradients et le snake n'est pas arrêté par le contour intérieur.

Pour ajuster les points des snakes, nous calculons le masque des pixels dents à l'intérieur de la bouche. La méthode utilisée pour segmenter les dents a été développée par Wang *et al.* [Wang, 2004b].
Les valeurs des composantes *u* et *a* des espaces *Luv* et *Lab* sont proches de 0 pour les pixels dents et plus élevées pour le reste des pixels de la bouche (cf. Section 4.3.1 et Fig. 4.14.e). En calculant les valeurs moyennes μ et les écart-types σ des composantes *u* et *a* des pixels de la bouche (sont pris en compte uniquement les pixels se trouvant dans la région définie par le contour extérieur des lèvres), un pixel (x, y) est défini comme un pixel « dent » si :

$$a(x,y) \le \mu_a - \sigma_a \quad \text{or} \quad u(x,y) \le \mu_u - \sigma_u \tag{5.03}$$

$a(x, y)$ et $u(x, y)$ sont les valeurs des composantes *u* et *a* du pixel (x, y). (μ_a, σ_a) et (μ_u, σ_u) sont les valeurs moyennes et les écart types calculés pour les pixels de la bouche.

La figure 5.23 montre des exemples du masque des pixels dents obtenu avec cette méthode (les pixels dents sont en verts).

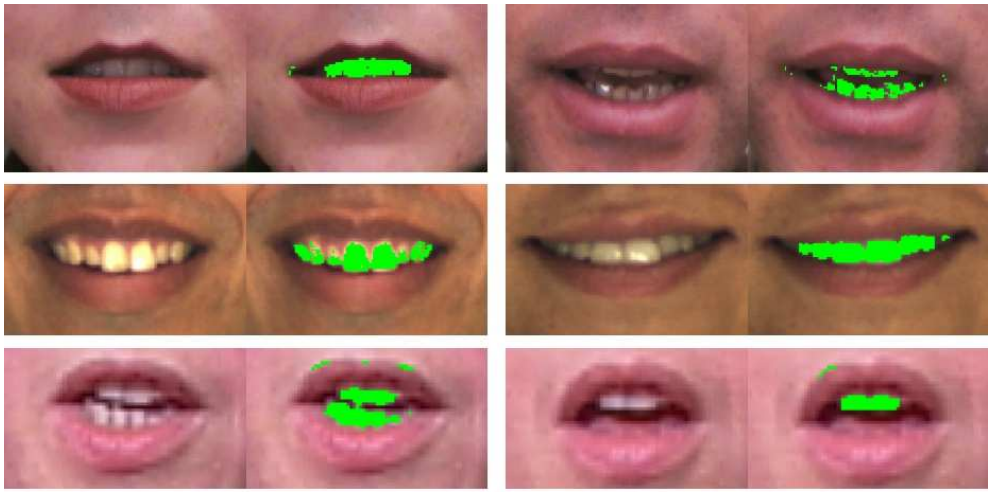


Fig. 5.23. Segmentation des dents.

A partir de ce masque, nous ajustons les points des snakes de la manière suivante :

- Si des pixels dents se trouvent au dessus du snake supérieur, les points du snake sont remontés jusqu'à qu'il n'y ait plus de pixels dents au dessus.
- Si des pixels dents se trouvent au dessous du snake inférieur, les points du snake sont abaissés jusqu'à qu'il n'y ait plus de pixels dents au dessous.

La figure 5.24 illustre l'ajustement du snake intérieur inférieur.

Fig. 5.24. Ajustement des snakes en présence des dents.

Nous pouvons utiliser le même principe d'ajustement pour les estimations P'_8 et P'_{10}, lorsque P'_8 (resp. P'_{10}) se trouve au dessous (resp. au dessus) des dents (cas de figure normalement impossible car ces points doivent être sur les contours intérieurs).

Des erreurs de convergence peuvent aussi arriver en présence des gencives. En effet, lorsque la couleur et la texture des gencives sont proches de celles des lèvres, le snake intérieur supérieur s'arrête sur la frontière séparant les gencives et les dents (cf. Fig. 5.25.a). Pour ajuster le snake dans ce cas, nous utilisons un second jumping snake pour le contour haut. Le germe de ce second snake est choisi comme étant le point clef P_8 trouvé grâce à la convergence du premier snake (point qui se trouve donc en dessous du vrai contour). Ainsi, les paramètres du second snake sont réglés de façon à ce que le snake se propage en dessus du germe (N_{up}=10 > N_{low}=3). Le gradient utilisé pour l'énergie externe du snake est le gradient G_5 défini par :

$$G_5(x,y)=\nabla[+Cr_N(x,y)+R_N(x,y)] \quad (5.04)$$

G_5 a été construit en considérant que les valeurs des composantes couleurs *Cr*, issue de l'espace *YCbCr*, et *R* sont plus faibles pour la lèvre supérieure que pour les gencives situées en dessous (d'où les signes +).

A la fin de la convergence, si le germe final du second snake est au dessous du point clef P_3 du contour extérieur supérieur, nous validons l'ajustement (cf. Fig. 5.25.b), sinon nous gardons le résultat du premier snake. En effet, dans le cas où il n'y a pas de gencives visibles, le second snake s'arrête au dessus de la bouche. La figure 5.26 montre des exemples de la convergence du second snake, lorsqu'il n'y pas de gencives visibles, pour des images des différentes bases utilisées dans cette thèse.

a. convergence du 1er snake

b. convergence du 2nd snake

Fig. 5.25. Ajustement du snake intérieur supérieur en présence des gencives.

a. Image de la base AR [Martinez, 1998]

b. Image de la base Gipsa

Fig. 5.26. Exemples de convergence du second snake lorsqu'il n'y a pas de gencives visibles.

5.3.2.b. Ajustement du modèle paramétrique intérieur « bouche ouverte »

Avec les différents ajustements des snakes, nous avons maintenant les 2 points clefs P_8 et P_{10} situés sur les contours intérieurs et plusieurs points utiles pour l'optimisation du modèle « bouche ouverte ». Il nous reste donc à trouver les commissures internes P_7 et P_9, et à ajuster le modèle. De la même façon que ce que nous avons adopté pour la détection des commissures P_1 et P_5 du contour extérieur, nous réalisons ces 2 étapes en une seule et même opération.

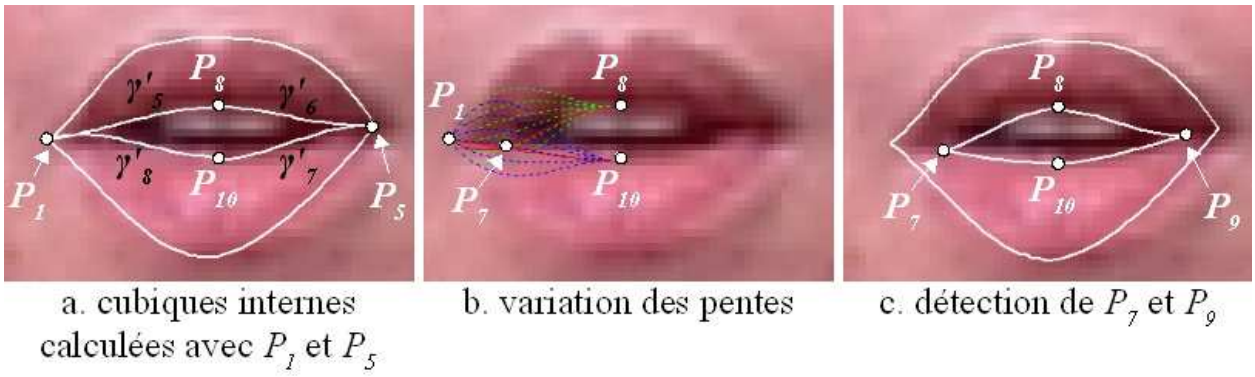
Dans un premier temps, nous supposons que les commissures internes sont les mêmes que les commissures externes ($P_7 = P_1$, et $P_9 = P_5$). Cette supposition est souvent vérifiée, notamment lorsque la bouche est grand ouverte. A partir de là, nous allons calculer des estimations des 4 courbes cubiques du modèle (notées : γ'_5, γ'_6, γ'_7 et γ'_8). Nous les appelons des estimations car elles sont obtenues en utilisant les commissures externes et non avec les bonnes commissures internes P_7 et P_9,. Nous connaissons les points extrêmes de ces 4 cubiques, à savoir :

- P_1 et P_8 pour γ'_5
- P_8 et P_5 pour γ'_6
- P_1 et P_{10} pour γ'_8
- P_{10} et P_5 pour γ'_7

De plus, nous imposons aussi comme contraintes une dérivée nulle en P_8 et P_{10}.

Ainsi pour chaque cubique, nous avons déjà 3 équations. En utilisant, les points des snakes proches de P_8 et P_{10}, une cubique est rapidement calculée pour chacun des 4 cas en utilisant la méthode des moindres carrés (cf. Fig. 5.27.a).

Maintenant, nous illustrons la stratégie choisie pour trouver la commissures P_7 (la même stratégie est adoptée pour détecter P_9). A partir des estimations γ'_5 et γ'_8, nous calculons plusieurs couples de cubiques en faisant varier les pentes des estimations entre P_1 and P_8 (γ_5) en haut, et entre P_1 and P_{10} (γ_8) en bas. Les pentes sont celles situées au niveau des points P_1 et P_5. 10 valeurs de pente autour de la valeur estimée sont testées pour chaque cubique. Le couple de cubiques qui maximise le flux moyen de gradient G_3 (pour γ_5) ou G_4 (pour γ_8) est gardé (les courbes rouges de la figure 5.27.b). La commissure P_7 est finalement choisie comme étant l'intersection de ces 2 cubiques. P_9 est trouvé de la même manière avec les cubiques γ_6 et γ_7.



a. cubiques internes calculées avec P_1 et P_5 | b. variation des pentes | c. détection de P_7 et P_9

Fig. 5.27. Détection des commissures internes.

La détection des commissures internes P_7 et P_9, et l'optimisation du modèle intérieur (bouche ouverte) sont bien effectuées en une seule et même opération.

Une évaluation de l'algorithme statique pour l'extraction du contour intérieur (cas bouche ouverte) des lèvres est présentée à la fin de ce chapitre.

5.3.2.c. Vérification de l'hypothèse « bouche ouverte »

Ayant obtenu le contour intérieur, il reste à vérifier a posteriori la validité de l'hypothèse « bouche ouverte » prise au début de la segmentation du contour intérieur. Si l'hypothèse est vérifiée, la recherche du contour intérieur des lèvres est terminée, sinon il faut passer au cas « bouche fermée » décrit dans la partie suivante.

La décision est prise à l'aide d'un critère géométrique sur la forme du contour intérieur trouvé. Après la convergence des snakes intérieurs, nous obtenons 3 cas de figures :

1) Si la bouche était en fait fermée, le contour intérieur est une ligne sombre séparant les lèvres supérieure et inférieure. Si cette ligne sombre n'est pas assez marquée, les snakes ne sont pas stoppés par le contour intérieur et le snake supérieur s'arrêtera en dessous du snake inférieur (cf. Fig. 5.28.a). Nous en déduisons que la bouche était fermée.

2) Si la bouche était fermée et que la ligne sombre est plus marquée, les 2 snakes s'arrêtent sur cette ligne et la distance entre les snakes est très faible (cf. Fig. 5.28.b). Si la surface de la région définie entre les 2 snakes est inférieure à un certain seuil, nous en déduisons que la bouche était fermée. Le seuil a été choisi expérimentalement et fixé à 10 pixels.

3) Si la bouche est ouverte, le snake supérieur s'arrête au dessus du snake inférieur, et la région définie entre les 2 snakes est plus grande que le seuil de 10 pixels (cf. Fig. 5.28.c).

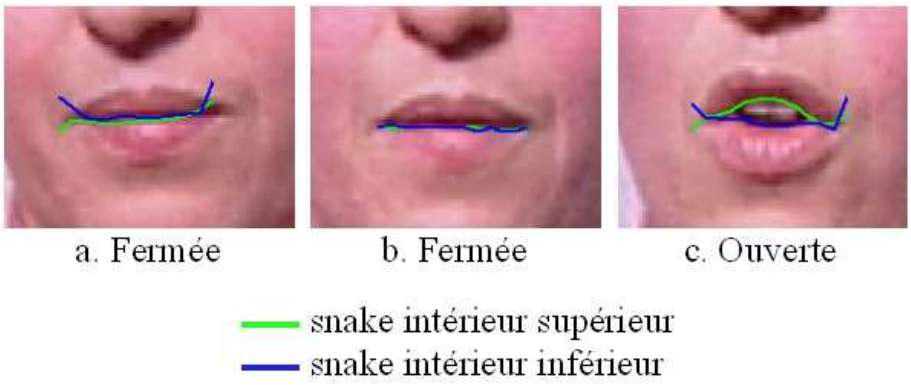


Fig. 5.28. Vérification de l'hypothèse bouche ouverte.

Les performances de la détection de l'état de la bouche sont évaluées dans le chapitre 6. La méthode de détection a été développée pendant l'étude sur le suivi des contours des lèvres (que nous présentons dans le chapitre suivant). Au commencement de nos travaux, nous supposions connue l'information sur l'ouverture de la bouche.

5.3.3. Segmentation du contour intérieur : cas d'une bouche fermée

Dans le cas où la bouche est fermée, le modèle paramétrique représentant le contour intérieur est composé de 2 courbes cubiques (cf. Fig. 5.29).

Dans cette partie, nous expliquons la stratégie adoptée pour détecter le point clef P_{11} (cf. Partie 5.3.3.a) et la méthode d'ajustement du modèle (cf. Partie 5.3.3.b).

5.3.3.a. Détection du point clef P_{11}

Lorsque la bouche est fermée, le contour intérieur est constitué de pixels lèvres et il peut être vu comme une ligne sombre reliant les 2 commissures P_1 et P_5 de la bouche. Pour initialiser la recherche du contour, nous utilisons la ligne des minima de luminance L_{min} introduite dans la partie 5.2.2. Comme nous l'avons déjà vu, la ligne L_{min} relie les pixels les plus sombres de l'intérieur de la bouche et elle est initialisée sur le pixel du segment $[P_3P_6]$ ayant la luminance la plus faible. En conséquence, comme nous pouvons le voir sur la figure 5.30.a, L_{min} est, en général, déjà une bonne représentation du contour intérieur. De plus, lors de l'extraction du contour extérieur, les commissures P_1 et P_5 ont été choisies sur cette ligne.
Il devient évident que le point clef P_{11} le plus adapté est le point initial de L_{min} (le pixel le plus sombre du segment $[P_3P_6]$).

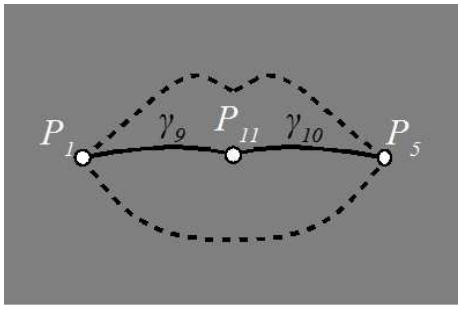


Fig. 5.29. Rappel : modèle paramétrique intérieur : bouche fermée.

5.3.3.b. Ajustement du modèle paramétrique intérieur « bouche fermée »

Dans le cas d'une bouche fermée, les commissures internes sont supposées être les mêmes que les commissures externes ($P_7 = P_1$, et $P_9 = P_5$). Aussi, en échantillonnant L_{min}, nous obtenons plusieurs points sur le contour intérieur. Pour chacune des 2 cubiques γ_9 et γ_{10}, nous avons donc les 2 points extrêmes et une condition d'annulation de leur dérivée en P_9. En utilisant les points issus de l'échantillonnage de L_{min} proches de P_9, les cubiques sont rapidement calculées en utilisant la méthode des moindres carrés. Dans le cas où la ligne des minima de luminance ne serait pas exactement sur le contour intérieur, la segmentation n'est pas assez précise. Une dernière étape d'optimisation consiste à faire varier les pentes des cubiques au niveau des commissures. Une dizaine de pentes autour des valeurs initiales sont testées pour chaque cubique et les courbes maximisant le flux moyen du gradient intensité (le contour intérieur étant une ligne sombre) sont choisies comme cubiques finales du modèle (les cubiques rouges de la figure 5.30.b).

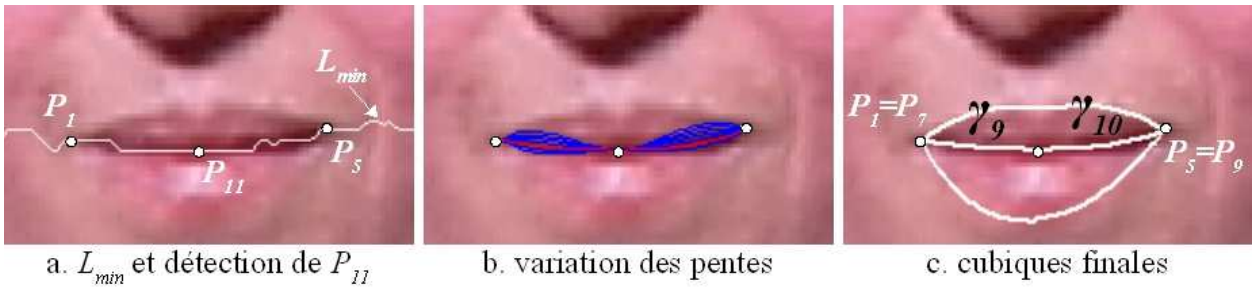


Fig. 5.30. Segmentation du contour intérieur : bouche fermée.

5.4. Evaluation quantitative des performances de l'algorithme statique

L'évaluation d'un algorithme de segmentation de contours est un problème difficile et il n'existe pas dans la littérature une méthode qui fasse l'unanimité. Nous pouvons distinguer deux manières de tester les performances de segmentation :

- soit en effectuant une évaluation quantitative; ce qui revient souvent à comparer les résultats avec une vérité de terrain,

- soit en évaluant la méthode au regard de l'application visée.

Dans cette section, nous allons présenter les résultats d'une évaluation quantitative pour l'algorithme statique. L'évaluation applicative sera étudiée plus en détails dans le chapitre 7.

Dans la section 5.4.1, nous présentons les données « vérité de terrain » créées pour l'évaluation quantitative des contours extérieur et intérieur des lèvres. La méthode d'évaluation utilisée pour comparer les résultats de l'expérimentation avec la vérité de terrain est illustrée dans la section 5.4.2. Finalement, les sections 5.4.3 et 5.4.4 montrent les résultats de l'évaluation pour le contour extérieur et pour le contour intérieur.

5.4.1. Construction de la vérité de terrain

Dans [Rehman, 2007], Rehman *et al.* étudient l'évaluation des algorithmes d'extraction du contour extérieur des lèvres à partir d'une vérité de terrain. Les auteurs montrent que malgré les avancées réalisées ces dernières années pour augmenter la précision de ces algorithmes, il n'existe pas d'études significatives qui comparent les performances des techniques proposées. Ceci s'explique par le fait qu'il n'existe pas une méthode d'évaluation standard pour évaluer la segmentation des contours des lèvres. Rehman *et al.* expliquent que la technique la plus communément employée consiste à comparer les résultats de segmentation avec des contours annotés manuellement par un expert humain. Ils montrent également que cette méthode d'évaluation n'est pas toujours efficace, car premièrement l'humain n'est pas forcément meilleur que la machine pour détecter des points dans l'image et deuxièmement l'expert humain introduit un bruit subjectif qui est trop important pour utiliser l'annotation manuelle directement comme vérité de terrain. En effet, pour une même image, différents experts humains produisent différentes annotations (cf. Fig. 5.31), de même qu'une seule personne peut fournir des annotations très variables d'un même contour.

Fig. 5.31. Variation de l'annotation manuelle du contour des lèvres pour 4 experts humains différents [Rehman, 2007].

Ainsi pour que la comparaison avec la vérité de terrain soit efficace, il faudrait pouvoir évaluer la précision de l'expert humain pour pouvoir modifier la vérité de terrain si besoin (ce qui est étudié par Rehman *et al.* en utilisant l'algorithme Expectation-Maximization) ou alors utiliser une base de données étiquetées par plusieurs experts humains (et faire une moyenne des données par exemple).

Dans cette thèse, nous n'avons pas construit de base de données de ce type. Les annotations ont été faites par un seul expert. L'évaluation quantitative de l'algorithme statique permet d'avoir une vue objective des performances et une indication sur la précision de la segmentation.

Pour construire la vérité de terrain, nous utilisons les images de la base Gipsa et de la base AR. (Les séquences de la base TELMA serviront à évaluer l'algorithme de suivi introduit dans le chapitre 6).

Nous avons annoté manuellement les contours extérieur et intérieur des lèvres pour des images où la bouche est ouverte. Ce qui correspond à 94 images de la base Gipsa et 507 images de la base AR [Martinez, 1998] (cf. Section 4.4). L'étiquetage manuel consiste à positionner plusieurs points sur les contours et à les relier par des segments de droite. La distance entre les points n'est pas obligatoirement constante (par exemple il y a plus de points au niveau de l'arc de Cupidon afin d'obtenir une plus grande précision) et typiquement l'annotation des contours labiaux nécessite le placement de quelques dizaines de points. Le positionnement des points a été effectué à l'aide du logiciel Matlab et en moyenne, il faut environ 2 minutes pour annoter une image. Ainsi l'étiquetage des 601 images a pris en tout environ 20 heures. Cependant, cette tâche demande une grande concentration et pour éviter des erreurs dues à la fatigue (lorsqu'un point était mal placé, il fallait reprendre l'annotation de l'image depuis le début), le travail a été fait en plusieurs fois et en plusieurs jours (la durée d'une séance d'étiquetage étant inférieure à 2 heures).

Pour le contour extérieur des lèvres, la méthode de segmentation est une amélioration de l'algorithme proposé dans [Eveno, 2003]. Ces améliorations concernent essentiellement l'automatisation de l'algorithme (calcul automatique des germes et réglage automatique des paramètres des jumping snakes) et l'utilisation des gradients introduits dans la section 4.3. Les performances sont très proches de l'algorithme original et une évaluation a déjà été effectuée dans [Eveno, 2003]. Nous avons donc seulement annoté manuellement les contours extérieurs des 94 images de la base Gipsa pour réaliser l'évaluation quantitative de la section 5.4.3. En revanche, nous avons annoté les contours intérieurs des lèvres des 94 images de la base Gipsa et des 507 images de la base AR pour l'évaluation de la section 5.4.4.

5.4.2. Méthode de comparaison

Pour comparer l'expérimentation avec la vérité de terrain, nous utilisons une méthode introduite par Wu *et al.* [Wu, 2002].

A partir de la vérité de terrain et des résultats expérimentaux, deux zones sont construites. La zone VT est la zone définie par les contours de la vérité de terrain (cf. Fig. 5.32.b). Une zone Algo est définie à partir des contours obtenus par l'algorithme de segmentation proposé (cf. Fig. 5.32.c).
Pour mesurer la différence entre les zones VT et Algo, nous procédons de la manière suivante :

- Si un pixel appartient à seulement une des deux zones à comparer, le pixel est choisi comme étant un pixel « erreur » (cf. Fig. 5.32.d). La figure 5.32 illustre la détermination des pixels « erreur » pour l'évaluation du contour intérieur.

Le taux d'erreur σ_e est défini comme étant le nombre de pixels « erreur », noté N_e, divisé par le nombre total de pixels à l'intérieur de la zone VT, noté N_{VT} :

$$\sigma_e = \frac{N_e}{N_{VT}} \tag{5.05}$$

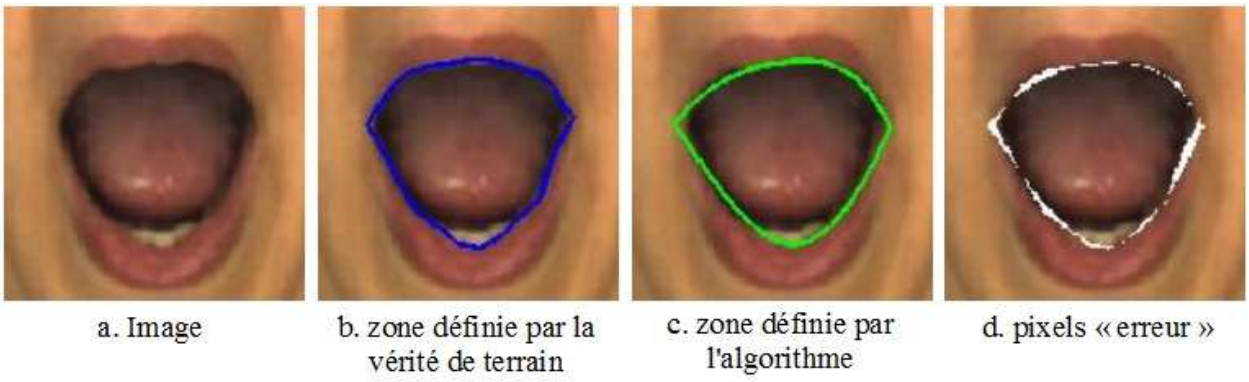


Fig. 5.32. Méthode de comparaison des résultats (exemple du contour intérieur).

5.4.3. Evaluation quantitative des performances pour le contour extérieur

Le tableau 5.01 montre le taux d'erreur moyen σ_e calculé en comparant les résultats de l'algorithme avec la vérité de terrain pour 94 images de la base Gipsa (images où la bouche est ouverte).

Taux d'erreur moyen (σ_e) en % (écart-type)	9,2 (2,5)
Nombre d'images avec $\sigma_e < 15\%$	92
Nombre d'images avec $15\% \leq \sigma_e < 25\%$	2
Nombre d'images avec $\sigma_e \geq 25\%$	0
Nombre moyen de pixels erreurs (N_e) (écart-type)	243 (83)
Nombre moyen de pixels dans la zone VT (N_{VT}) (écart-type)	2631 (485)

Tab. 5.01. Évaluation quantitative du contour extérieur pour 94 images de la base Gipsa.

Le taux d'erreur moyen σ_e obtenu avec les 94 images est de 9,2% et la quasi-totalité des images donne un taux d'erreur inférieur à 15% (92 images). Le tableau 5.01 donne également le nombre moyen de pixels erreurs et le nombre moyen de pixels de la zone VT pour une image de la base Gipsa bouche ouverte.

La figure 5.33 montre quelques exemples de segmentation du contour extérieur représentatifs de la précision de l'algorithme proposé pour des images de la base Gipsa. Les exemples concernent à la fois des cas « bouche ouverte » et des cas « bouche fermée ». Les figures 5.34 à 5.36 montrent respectivement des exemples avec la base AR pour des bouches fermées, pour des sourires (caractéristique 2 et 15, cf. Section 4.4.2) et pour des cris (caractéristique 4 et 17, cf. Section 4.4.2). Les résultats sont centrés sur la bouche pour avoir une meilleure visualisation de la segmentation.

Fig. 5.33. Exemples de segmentation du contour extérieur pour des images de la base Gipsa.

Fig. 5.34. Exemples de segmentation du contour extérieur pour des images de la base AR [Martinez, 1998] où la bouche est fermée.

Les images variées que nous avons à disposition avec les bases Gipsa et AR permettent d'illustrer la flexibilité du modèle paramétrique du contour extérieur des lèvres composé de quatre cubiques et d'une ligne brisée (pour modéliser l'arc de Cupidon). En effet, le modèle peut suivre le contour des lèvres pour des bouches fermées (cf. Fig. 5.33 et 5.34), ouvertes (cf. Fig. 5.33 et 5.35) ou largement ouvertes (cf. Fig. 5.36) quelque soient les déformations de la bouche que cela implique. Le modèle est également aussi bien adapté pour des lèvres fines que pour des lèvres plus charnues.

De plus, la segmentation est possible, même en cas de présence de moustaches ou de barbes (visibles sur plusieurs images des figures 5.33 à 5.36).

Fig. 5.35. Exemples de segmentation du contour extérieur pour des images de la base AR [Martinez, 1998] (caractéristique « sourire »).

Fig. 5.36. Exemples de segmentation du contour extérieur pour des images de la base AR [Martinez, 1998] (caractéristique « cri »).

Des erreurs de segmentation peuvent arriver lorsque les contours des lèvres ne sont pas assez marqués et que la différence entre les pixels peau et les pixels lèvre n'est pas assez importante. Dans ce cas, les pseudo-teintes *H1* et *H2* (cf. Section 4.3) ne permettent pas de discriminer efficacement ces deux groupes de pixels. Les jumping snakes utilisés pour détecter le contour extérieur des lèvres ne sont pas freinés et ils dépassent le contour labial. Il en résulte que le modèle paramétrique extérieur optimisé par la position des points des snakes ne se trouve pas sur le contour. La figure 5.37 montrent des exemples de mauvaises segmentations pour le contour extérieur haut et/ou bas (images du milieu). La recherche du contour intérieur est également représentée, mais celui-ci sera discuté dans la section suivante.

Fig. 5.37. Exemples de mauvaises segmentations dues au contour labial peu marqué. Images de la base AR [Martinez, 1998].

La présence de moustaches peut cacher le contour extérieur haut des lèvres ou créer un contour parasite qui gène la croissance du jumping snake extérieur haut (cf. Fig. 5.38).

Fig. 5.38. Exemples de mauvaises segmentations en cas de présence de moustaches. Images de la base AR [Martinez, 1998].

Une autre difficulté rencontrée par notre méthode est la précision de la position des commissures des lèvres qui dans certains cas ne sont pas exactement à leur place. La figure 5.39 montre des exemples où une des commissures est soit trop à l'extérieur, soit trop à l'intérieur de la bouche.

Fig. 5.39. Exemples de mauvaises détections de la position des commissures.

Dans le cadre du logiciel de maquillage de Vesalis, nous avons remarqué que la segmentation du contour extérieur des lèvres peut être erronée lorsque la personne a la peau noire et que les lèvres ont une couleur proche de celle de la peau. Ceci est dû à deux raisons :

- les gradients développés en combinant des informations de luminance et de chrominance ne sont plus aussi performants pour accentuer le contour des lèvres,
- la ligne de minimum de luminance L_{min} ne passe plus forcément par les commissures (les parties sombres du visage étant plus nombreuses pour une personne ayant la peau noire, la ligne ne suit pas obligatoirement l'intérieur de la bouche).

La figure 5.40 montre des exemples où la méthode de segmentation fonctionne car la différence de couleur entre les pixels lèvre et les pixels peau est suffisamment importante. Les images sont issues de la base FERET [Phillips, 1998].

Fig. 5.40. Exemples de segmentation réussies pour des personnes ayant la peau noire.

a. Exemples de segmentations erronées

b. Amélioration de la segmentation

Fig. 5.41. Amélioration de la segmentation pour des personnes ayant la peau noire.

En revanche, la figure 5.41.a illustre la difficulté rencontré par l'algorithme proposé pour détecter le contour extérieur des lèvres lorsque la peau et les lèvres ont une couleur trop proche. Les images sont issues de la base FERET [Phillips, 1998]. Lors du développement de l'algorithme, ce problème n'a pas été mis en évidence car peu de personnes avec la peau noire étaient présentes dans les bases d'images

utilisées. En modifiant les gradients utilisés pour la convergence des jumping snakes extérieurs et pour l'optimisation du modèle paramétrique du contour extérieur, il est possible d'améliorer la segmentation (cf. Fig. 5.41). En outre, des erreurs de segmentation peuvent arriver car la ligne des minima de luminance L_{min} ne passe par les commissures (cf. Fig. 5.42.b). Dans ce cas, la détection peut être améliorée en utilisant une autre méthode pour calculer L_{min} (qui consiste seulement à tester les pixels les plus sombres au niveau de la bouche sans que ceux-ci soient chaînés, cf. Fig. 5.42.c).

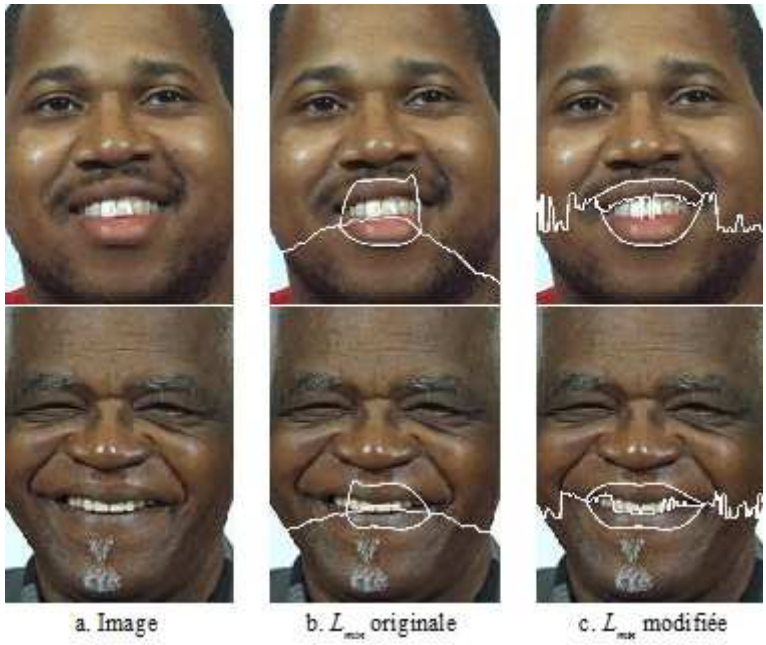


Fig. 5.42. Amélioration de la segmentation en modifiant le calcul de L_{min}.

Cette étude a été réalisée sur une soixantaine d'images issues de la base FERET [Phillips, 1998]. Pour être concluante, il faudrait réaliser la même étude sur un plus grand nombre d'image. Malheureusement, nous n'avons pas eu le temps de le faire car il aurait fallu fabriquer une base d'images (le nombre d'images de personnes ayant la peau noire dans étant très faible dans les bases d'images que nous avions à disposition). Cependant sur les images testées, nous avons constaté que soit la méthode originale fonctionne, soit la segmentation peut être améliorée par les traitements illustrés avec les figures 5.41 et 5.42. Ainsi, une perspective d'amélioration de l'algorithme serait d'envisager d'avoir deux méthodes de segmentation différentes, mais il faudrait pouvoir savoir laquelle choisir avant le traitement. Le fait de détecter la couleur de la peau du visage a traité n'est pas un critère suffisant, car chacune des deux méthodes peut donner de bons résultats pour une personne à la peau noire et de mauvais avec une autre.

5.4.4. Evaluation quantitative des performances pour le contour intérieur

Les tableaux 5.02, 5.03 et 5.04 montrent le taux d'erreur moyen σ_e calculé en comparant les résultats de l'algorithme avec la vérité de terrain pour, respectivement, les 94 images de la base Gipsa où la bouche est ouverte, les 252 images de la base AR avec la caractéristique « sourire » et les 255 images de la base AR avec la caractéristique « cri ». Le contour intérieur n'est évalué que dans le cas où la bouche est ouverte.

Le taux d'erreur moyen σ_e obtenu avec les 94 images est de 18,8% (avec un écart-type de 9,5%). Pour les images de la base AR, σ_e vaut 25,2% (avec un écart-type de 9,3%) pour la caractéristique « sourire » et 11,2% (avec un écart-type de 9,5%) pour la caractéristique « cri ». Le taux d'erreur moyen σ_e est plus faible pour la caractéristique « cri » que pour la caractéristique « sourire »; cette différence est due à la méthode de comparaison utilisée. En effet, le nombre de pixels erreurs N_e dans une image est relativement

constant pour l'ensemble de la base AR. Or pour calculer σ_e, N_e est divisé par le nombre de pixels N_{VT} présents dans la zone VT (cf. Eq. 5.05). Cependant le nombre moyen de pixels N_{VT} est plus important (3,3 fois plus grand) pour la caractéristique « cri » (4968 pixels en moyenne, cf. Tab. 5.04) que pour la caractéristique « sourire » (1505 pixels en moyenne, cf. Tab. 5.03), alors que le nombre moyen de pixels erreur N_e est seulement 1,4 fois plus grand pour la caractéristique « cri » que pour la caractéristique « sourire » (535 contre 360). En conséquence, le taux d'erreur est plus faible.

Base Gipsa (94 images de bouches ouvertes)	
Taux d'erreur moyen (σ_e) en % (écart-type)	18,8 (6,8)
Nombre d'images avec $\sigma_e < 15\%$	28
Nombre d'images avec $15\% \leq \sigma_e < 25\%$	49
Nombre d'images avec $25\% \leq \sigma_e < 50\%$	17
Nombre d'images avec $50\% \leq \sigma_e \leq 75\%$	0
Nombre d'images avec $\sigma_e > 75\%$	0
Nombre moyen de pixels erreurs (N_e) (écart-type)	108 (43)
Nombre moyen de pixels dans la zone VT (N_{VT}) (écart-type)	616 (238)

Tab. 5.02. Évaluation quantitative du contour intérieur pour 94 images de la base Gipsa.

Base AR (252 images, caractéristique « sourire »)	
Taux d'erreur moyen (σ_e) en % (écart-type)	25,2 (9,3)
Nombre d'images avec $\sigma_e < 15\%$	26
Nombre d'images avec $15\% \leq \sigma_e < 25\%$	118
Nombre d'images avec $25\% \leq \sigma_e < 50\%$	103
Nombre d'images avec $50\% \leq \sigma_e \leq 75\%$	5
Nombre d'images avec $\sigma_e > 75\%$	0
Nombre moyen de pixels erreurs (N_e) (écart-type)	360 (179)
Nombre moyen de pixels dans la zone VT (N_{VT}) (écart-type)	1505 (598)

Tab. 5.03. Évaluation quantitative du contour intérieur pour 252 images de la base AR [Martinez, 1998]. Caractéristique « sourire ».

Base AR (255 images, caractéristique « cri »)	
Taux d'erreur moyen (σ_e) en % (écart-type)	11,2 (9,5)
Nombre d'images avec $\sigma_e < 15\%$	216
Nombre d'images avec $15\% \leq \sigma_e < 25\%$	19
Nombre d'images avec $25\% \leq \sigma_e < 50\%$	16
Nombre d'images avec $50\% \leq \sigma_e \leq 75\%$	4
Nombre d'images avec $\sigma_e > 75\%$	0
Nombre moyen de pixels erreurs (N_e) (écart-type)	535 (497)
Nombre moyen de pixels dans la zone VT (N_{VT}) (écart-type)	4968 (1556)

Tab. 5.04. Évaluation quantitative du contour intérieur pour 255 images de la base AR [Martinez, 1998]. Caractéristique « cri ».

Fig. 5.43. Exemples de segmentation du contour intérieur pour des images de la base Gipsa (cas « bouche fermée »).

Fig. 5.44. Exemples de segmentation du contour intérieur pour des images de la base Gipsa (cas « bouche ouverte »).

Fig. 5.45. Exemples de segmentation du contour intérieur pour des images de la base AR où la bouche est fermée.

Fig. 5.46. Exemples de segmentation du contour intérieur pour des images de la base AR (caractéristique « sourire »).

Fig. 5.47. Exemples de segmentation du contour intérieur pour des images de la base AR (caractéristique « cri »).

La figure 5.43 montre des exemples de segmentation du contour intérieur lorsque la bouche est fermée et la figure 5.44 montre des exemples avec la bouche ouverte pour des images de la base Gipsa. Les figures 5.45 à 5.47 montrent respectivement des exemples avec la base AR pour des bouches fermées, pour des sourires et pour des cris. Les résultats sont centrés sur la bouche pour avoir une meilleure visualisation de la segmentation.

Les bases Gipsa et AR permettent d'illustrer l'efficacité de l'algorithme proposé pour segmenter le contour intérieur des lèvres, quelque soit l'ouverture de la bouche : fermée (cf. Fig. 5.43 et 5.45), ouverte (cf. Fig. 5.44 et 5.46) ou très ouverte (cf. Fig. 5.47). La caractéristique « sourire » de la base AR est très intéressante par rapport à l'application du logiciel de maquillage (Makeuponline), dans la mesure où cela représente parfaitement le cas pratique (généralement, une personne voulant se maquiller ferme la bouche ou sourit). Les exemples montrent que la segmentation est effectuée pour toutes les configurations internes possibles de la bouche (frontière Lèvre/Lèvre (bouche fermée), Lèvre/Dent (cf. Fig. 5.46), Lèvre/Cavité Orale (cf. Figure 5.46), Lèvre/Gencive (cf. Fig. 5.47) et Lèvre/Langue (cf. Fig. 5.47)).
On peut également remarquer que la présence de commissures internes apparaît seulement pour les images de la base Gipsa. Ceci s'explique par le fait que pour la base AR les bouches sont suffisamment ouvertes pour que les commissures externes et internes soient confondues. Les commissures internes apparaissant lorsque la bouche est légèrement ouverte, celles-ci sont plus visibles dans les séquences vidéo. Ainsi nous verrons dans le chapitre suivant, qui traite du suivi des contours des lèvres, que la détection de commissures internes se produit souvent dans les images de la base TELMA qui montre une personne en train de parler.

La majorité des erreurs de segmentation du contour intérieur se produisent à cause de la présence de la langue. Lorsque la frontière Lèvre inférieure/Langue n'est pas assez marquée, le jumping snake intérieur bas n'est pas stoppé par le contour intérieur des lèvres et il s'arrête sur la langue (cf. Fig. 5.48). Pour la base AR où la bouche est ouverte, 16,5% des images où la langue est visible et forme une frontière Lèvre inférieure/Langue, amènent une erreur de segmentation.

Fig. 5.48. Exemples d'erreurs de segmentation dues à la présence de la langue.

Ensuite, des erreurs de segmentation peuvent se produire pour trois raisons :

- Lorsque les gencives et les lèvres ont des caractéristiques couleurs trop proches (le contour extrait se retrouve au dessous du vrai contour labial interne, cf. Fig. 5.49). Dans ce cas, l'ajustement des snakes (cf. Section 5.3.2) ne permet pas de trouver le bon contour.
- Lorsque les dents sont sombres ou apparaissent jaunes (le contour extrait se retrouve sur les dents, cf. Fig. 5.50). L'ajustement des snakes ne permet pas de trouver le bon contour, car les dents ne sont pas brillantes et l'algorithme de segmentation des pixels dents n'est donc pas efficace (cf. Section 5.3.2).
- lorsque une partie de la lèvre inférieure est trop exposée à la lumière (le contour se bloque sur la région brillante de la lèvre inférieure, le snake étant réglé pour se bloquer sur les parties brillantes telles que les dents.cf. Fig. 5.51).

Fig. 5.49. Exemples d'erreurs de segmentation dues à la présence des gencives.

Fig. 5.50. Exemples d'erreurs de segmentation dues aux dents trop sombres ou jaunes.

Fig. 5.51. Exemples d'erreurs de segmentation dues à une région brillante sur la lèvre inférieure.

5.5. Conclusion

Dans ce chapitre, nous avons proposé un algorithme automatique de détection des contours extérieur et intérieur des lèvres pour des images statiques.

La méthode est une combinaison de contours actifs et de modèles paramétriques composés de cubiques.
Le type de contour actif utilisé est le jumping snake [Eveno, 2003], qui permet une initialisation simple (positionnement d'un germe) et relativement loin du contour à segmenter. La convergence de quatre jumping snakes (deux snakes extérieurs haut et bas, et deux snakes intérieur haut et bas) permet de trouver la position de plusieurs points clefs externes et internes, et de positionner les modèles paramétriques. Un modèle est défini pour le cas « bouche ouverte » et un autre pour le cas « bouche fermée ». L'optimisation des courbes des modèles et la détection des commissures (externes et internes) sont réalisées en une seule et même opération en maximisant le flux moyen de plusieurs gradients développés spécifiquement pour chaque partie du contour labial (contour extérieur haut et bas, et contour intérieur haut et bas). Les gradients ont été construits à partir de la luminance et de plusieurs composantes couleurs; en particulier de deux pseudo-teintes, qui accentuent le contraste entre les pixels lèvre et les pixels peau.

Nous avons également effectué une évaluation quantitative des performances de l'algorithme statique, qui consiste à comparer les résultats expérimentaux avec de la vérité de terrain. L'évaluation montre l'efficacité de la méthode proposée et la flexibilité des modèles paramétriques, qui peuvent modéliser précisément les déformations de la bouche.

CHAPITRE 6

Suivi des contours des lèvres

Après avoir développé un algorithme de segmentation pour des images statiques (cf. Chapitre 5), la stratégie pour le suivi peut prendre deux directions distinctes :

1) soit nous appliquons l'algorithme statique sur chaque image de la séquence; ceci revient à traiter les images séparément comme un assemblage d'images statiques,
2) soit nous développons un algorithme de suivi qui utilise des informations temporelles afin de prendre en compte l'enchaînement des images de la séquence.

Nous avons choisi de créer un module de suivi pour améliorer les performances de la segmentation en termes de vitesse et de robustesse. En effet, par rapport à un algorithme statique qui traiterait une nouvelle image de la séquence sans connaissances *a priori*, un algorithme de suivi utilise des informations temporelles définies à partir des images précédentes. Le traitement d'une nouvelle image tient compte ainsi des informations sur l'emplacement de la bouche ou sur la forme de la bouche (épaisseur des lèvres, allure des courbes des modèles dans les anciennes images...). Ces données permettent une segmentation plus rapide, mais aussi plus robuste, car les résultats précédents peuvent servir aux cas mal résolus en statique (exemple de la présence de la langue, cf. Section 5.4.4), à condition que la détection dans l'image précédente soit exacte.

La figure 6.01 montre le schéma global de l'algorithme de suivi et l'enchaînement des différents modules utilisés.

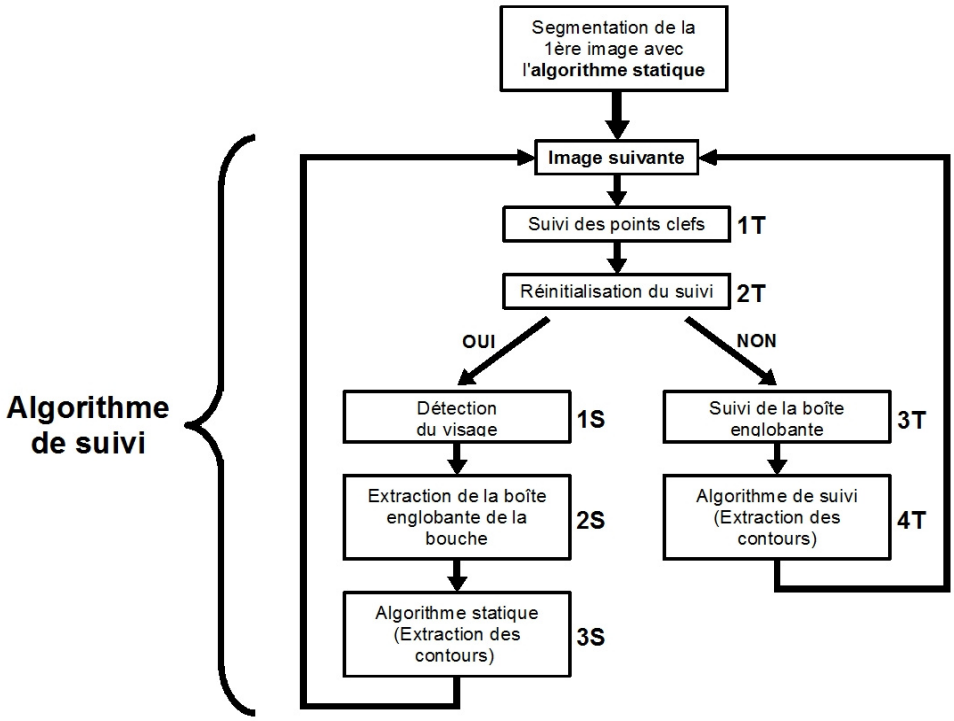


Fig. 6.01. Synoptique de l'algorithme de suivi des contours des lèvres.

Dans ce chapitre, nous présentons les 4 phases de la méthode de suivi des contours de la bouche dans une séquence vidéo. On suppose que le contour de la bouche a été extrait dans la première image. Pour les images suivantes, nous utilisons les mêmes modèles paramétriques que ceux présentés dans la partie 4.1.

Pour les séquences vidéo où le visage est vu en entier, la région d'intérêt est extraite avec l'algorithme CFF (cf. Partie 5.1.1). Cependant, la détection du visage est réalisée uniquement dans la 1ère image et les mêmes coordonnées du cadre du visage sont utilisées pour le reste de la séquence. Un exemple est présenté sur la figure 6.02. Une nouvelle séquence est construite à partir des régions d'intérêt de chaque image de l'ancienne séquence. En conséquence, le visage peut bouger au cours de la vidéo, mais il est impératif que la bouche reste dans le cadre initial du visage (ceci est vérifiée pour l'ensemble des séquences de la base TELMA (cf. Section 4.4.3).
Le fait de garder les mêmes coordonnées pour le cadre du visage pour l'ensemble du corpus permet pour le module de suivi de conserver les informations obtenues dans les images précédentes (coordonnées des points clefs, coordonnées des contours, position de la boîte autour de la bouche ou valeurs des paramètres des cubiques) sans changer de repère; les sous-images de la nouvelle séquence faisant toutes la même taille. Pour des séquences où la tête aurait des mouvements plus larges, il faudrait ajouter une étape de suivi du cadre du visage. Il serait envisageable, par exemple, d'utiliser un filtre de Kalman, comme nous le faisons pour suivre la boîte autour de la bouche (cf. Partie 6.3).

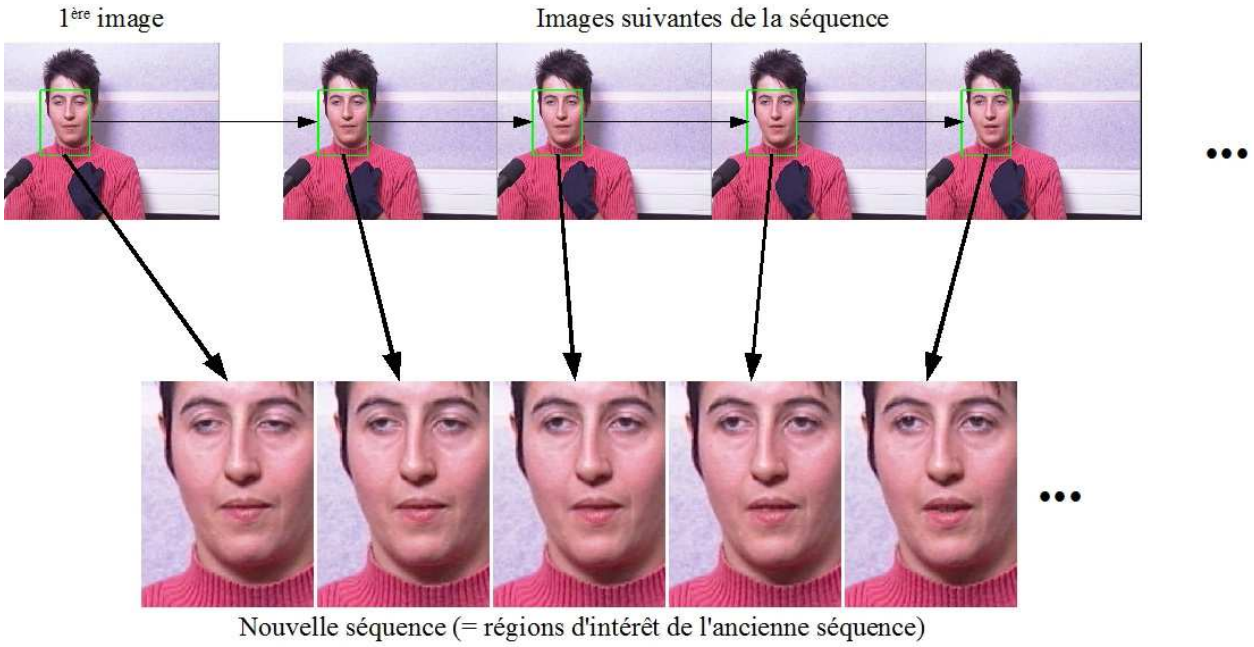


Fig. 6.02. Extraction de la zone d'intérêt de chaque image du corpus avec les coordonnées de la boîte englobante du visage obtenue à partir de la 1ère image de la séquence.

La partie 6.1 détaille la phase 1T de la méthode de suivi. Cette étape consiste à suivre les points clefs externes et internes des modèles paramétriques, d'une image à la suivante, en utilisant la méthode de Lucas-Kanade. Le positionnement de ces points clefs étant un facteur déterminant pour obtenir des contours des lèvres précis, il est nécessaire de procéder à un ajustement des estimations fournies par l'algorithme de Lucas-Kanade.

Dans la partie 6.2, nous expliquons comment, à partir des résultats du suivi des points clefs, nous déterminons si le suivi des contours des lèvres doit être réinitialisé à cause d'une mauvaise détection. En cas de réinitialisation, nous recommençons la segmentation des contours des lèvres, en utilisant l'algorithme statique décrit dans le chapitre 5, sur l'image de la séquence qui pose problème. Si le suivi des points clefs est satisfaisant, nous continuons le processus de l'algorithme de suivi des contours des lèvres.

Dans la partie 6.3, nous présentons le filtre de Kalman utilisé pour suivre la boîte autour de la bouche (phase 3T). Ceci permet notamment de limiter la recherche des courbes, représentant les contours extérieur et intérieur, à la région définie par la boîte.

La méthode d'extraction des contours extérieur et intérieur des lèvres, dans le cadre du suivi, est détaillée dans la partie 6.4. Les positions initiales des courbes des modèles paramétriques sont déterminées à partir du résultat de la segmentation dans l'image précédente. L'utilisation de l'information temporelle permet de rendre la détection des contours plus robuste. Les positions finales des courbes sont obtenues avec la méthode de maximisation de flux moyen de gradient utilisée dans l'algorithme statique (cf. Chapitre 5).

Finalement, une évaluation quantitative des performances est réalisée dans la partie 6.5 en comparant des contours obtenus avec l'algorithme de suivi avec ceux de la vérité de terrain.

6.1. Suivi des points clefs des modèles paramétriques (Phase 1T)

Les modèles paramétriques, que nous utilisons pour modéliser les contours extérieur et intérieur des lèvres (cf. Section 4.1), sont positionnés à partir des différents points clefs externes et internes $P_{i=1 \text{ à } 11}$. Ces points ont été trouvés sur la première image avec l'algorithme statique, en faisant converger différents jumping snakes. Pour les images suivantes de la séquence, nous déterminons la position des points clefs en effectuant un suivi de ces points image par image. La détection des points est, de ce fait, plus rapide et plus robuste que pour la méthode statique (cf. Chapitre 5).

Le suivi des points doit être le plus précis possible dans la mesure où ils vont permettre d'initialiser la position des modèles paramétriques dans l'image courante. Le résultat de la segmentation des lèvres est donc étroitement lié à la position des points clefs. Dans [Baron, 1994], Baron *et al.* ont montré que les méthodes d'estimation de mouvement les plus précises sont la méthode différentielle du premier ordre de Lucas et Kanade [Lucas, 1984] et la méthode de phase de Fleet et Jepson [Fleet, 1990]. Nous avons choisi l'algorithme de Lucas-Kanade (méthode largement utilisée dans la littérature pour le suivi de points), car celui-ci est beaucoup plus rapide que celui de Fleet-Jepson, qui implique de réaliser plusieurs filtrages.

6.1.1. L'algorithme de Lucas-Kanade

L'algorithme, que nous utilisons dans ce travail, est une variante de la méthode originale d'estimation du flux optique développée par Lucas et Kanade [Lucas, 1984]. La théorie présentée dans cette section provient de l'étude proposée par Tomasi et Kanade [Tomasi, 1991].

Une séquence d'images est représentée par la fonction $I(x, y, t)$ où x et y sont les variables d'espace et t la variable temporelle. Dans notre cas, I est la valeur de la luminance du pixel (x, y) à l'instant t. Si les images de la séquence sont acquises à une cadence suffisamment élevée, nous pouvons supposer qu'un pixel (x, y) de l'image courante se retrouve dans l'image suivante par une translation.
Cette corrélation peut s'exprimer de la manière suivante :

$$I(x, y, t+1) = I(x-\xi, y-\eta, t) \tag{6.01}$$

La translation du pixel $X=(x, y)$ de l'image courante (à l'instant t) à l'image suivante (à l'instant t+1) est représentée dans l'équation 6.01 par le vecteur déplacement $d=(\xi, \eta)$.
L'équation 6.01 peut se récrire plus simplement en enlevant la variable temporelle et en définissant $J(X)=I(x, y, t+1)$ et $I(X-d)=I(x-\xi, y-\eta, t)$:

$$J(X) = I(X - d(X)) + n(X) \tag{6.02}$$

où $n(X)$ est le bruit.

Il est difficile de suivre un point d'une image à la suivante en ne considérant que le pixel en lui-même, à moins d'avoir une valeur de luminance très différente de celle des pixels voisins. Considérant une région W de taille mxm autour du point que l'on souhaite suivre dans l'image de référence I, l'objectif est de déterminer la région de même taille la plus ressemblante dans l'image suivante J. Cette région a été translatée du vecteur d de l'instant t à l'instant t+1. Le vecteur de déplacement d est évalué en minimisant l'erreur ε calculée sur le voisinage W :

$$\varepsilon = \sum_{X \in W} [I(X - d(X)) - J(X)]^2 \omega(X)\, dX \tag{6.03}$$

où $\omega(X)$ est une fenêtre de pondération. Pour la suite, nous prenons $\omega(X) = 1$ (valeur utilisée dans le cadre de cette thèse).

Le déplacement qui minimise l'erreur ε est déterminé de manière itérative. Soit $d^i(X)$ le déplacement à l'itération i, le déplacement à l'itération suivante est :

$$d^{i+1}(X) = d^i(X) + \Delta d^i(X) \tag{6.04}$$

où $\Delta d^i(X)$ est le déplacement incrémental que l'on souhaite calculer. En considérant que la valeur du déplacement est la même pour tous les pixels de la région W (cf. Fig. 6.03), l'équation 6.04 devient :

$$d^{i+1} = d^i + \Delta d^i \tag{6.05}$$

L'équation 6.02 peut se récrire de la manière suivante :

$$\begin{aligned} J(X) &= I(X - d^{i+1}) \\ &= I(X - (d^i + \Delta d^i)) \end{aligned} \tag{6.06}$$

L'équation précédente peut être approchée avec un développement de Taylor au premier ordre :

$$I(X - d^{i+1}) \approx I(X - d^i) - g^T \Delta d^i \tag{6.07}$$

où g est le vecteur gradient :

$$g^T = \left(\frac{\partial I(X)}{\partial x} \quad \frac{\partial I(X)}{\partial y} \right) \tag{6.08}$$

L'équation 6.03 définissant l'erreur ε à minimiser devient :

$$\begin{aligned} \varepsilon &\approx \sum_{X \in W} [I(X - d^i) - g^T \Delta d^i - J(X)]^2 \\ &= \sum_{X \in W} [h - g^T \Delta d^i]^2 \end{aligned} \tag{6.09}$$

avec $h = I(X - d^i) - J(X)$.

Pour obtenir, la valeur qui minimise l'erreur ε, on dérive l'équation 6.09 par rapport à Δd^i :

$$\begin{aligned} \frac{\partial \varepsilon}{\partial \Delta d^i} &= 2 \sum_{X \in W} (h - g^T \Delta d^i) g \\ &= 2 \left(\sum_{X \in W} hg \right) - 2 \left(\sum_{X \in W} g g^T \right) \Delta d^i \end{aligned} \tag{6.10}$$

L'annulation de cette dérivée fournit la relation fondamentale de l'algorithme de Lucas-Kanade :

$$G \Delta d^i = e \tag{6.11}$$

où :

$$G = \sum_{X \in W} g g^T$$

$$e = \sum_{X \in W} (I(X - d^i) - J(X)) g$$

Pour les conditions initiales, on fixe $d^0=[0\ 0]^T$. Ensuite, le déplacement incrémental est déterminé à l'aide de l'équation 6.11. Le processus d'itération se termine lorsque le déplacement incrémental devient très faible ou après un nombre d'itérations fixé.
La matrice *G* est obtenue en calculant le gradient de la luminance sur la fenêtre de référence *W*.

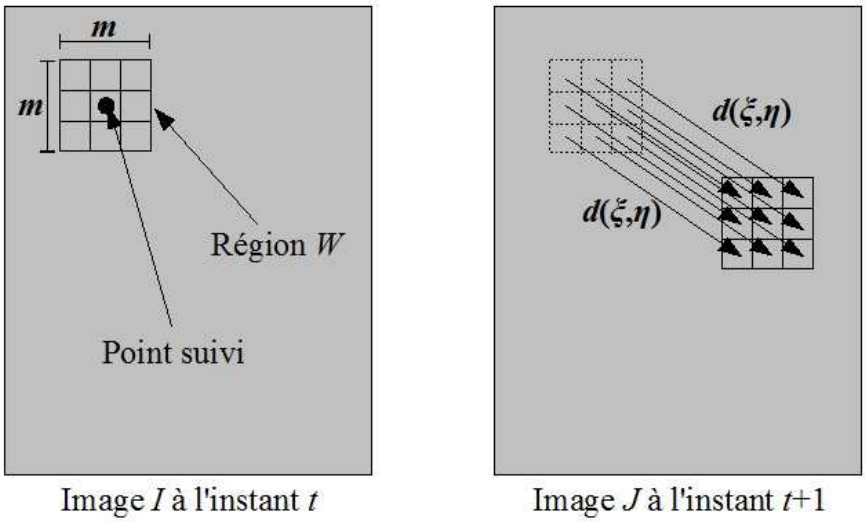


Fig. 6.03. Voisinage du point suivi retrouvé dans l'image suivante par une translation.

6.1.2. L'algorithme de Lucas-Kanade appliqué au suivi des points clefs externes et internes

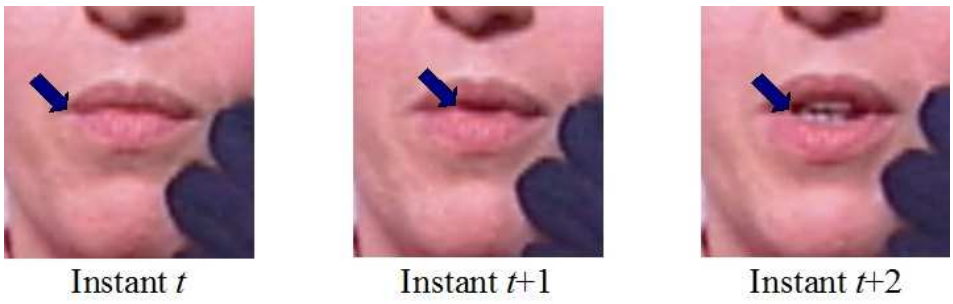
Dans le contexte du suivi des contours extérieur et intérieur des lèvres, nous utilisons l'algorithme de Lucas-Kanade pour suivre 7 des 11 points clefs externes et internes des modèles paramétriques. Les points P_1, P_2, P_4, P_5, P_6, P_8 et P_{10} sont suivis à l'aide de cette technique.

Connaissant les positions des points P_2 et P_4, nous verrons que le point P_3 est trouvé rapidement en testant des pixels sur la colonne médiatrice du segment $[P_2P_4]$ (cf. Section 6.1.3.b). Les positions des commissures internes P_7 et P_9 sont calculées directement pour toute nouvelle image de la séquence (cf. Section 6.4.2). Ce choix a été fait dans la mesure où ce sont deux points difficiles à suivre. En effet, le voisinage des commissures internes peut changer significativement d'une image à l'autre car :

1) les commissures internes se déplacent rapidement quand la bouche s'ouvre ou se ferme; la figure 6.04.a montre un exemple où la commissure intérieure gauche se déplace rapidement vers le centre de la bouche lorsque la bouche passe de l'état fermé à l'état ouvert,
2) l'intérieur de la bouche change brutalement d'apparence lorsque les dents, la langue, les gencives ou la cavité orale, apparaissent ou disparaissent. Sur la figure 6.04.b, la commissure intérieure gauche n'est entourée que de lèvre et de peau à l'instant *t*, puis il y a apparition de la cavité orale et des dents sur les deux images suivantes.

De ce fait, l'algorithme de suivi de Lucas-Kanade ne peut pas être performant pour suivre P_7 et P_9.

Enfin, nous avons vu, dans le chapitre 5, que le cas « bouche fermée » est un cas simple et qu'il conduit à une segmentation rapide du contour intérieur. Lorsque la bouche est détectée fermée, le point clef interne P_{11} est déterminé de la même manière que pour l'algorithme statique (cf. Section 6.4.2).



a. Exemple de déplacement rapide de la commissure interne gauche lorsque la bouche s'ouvre

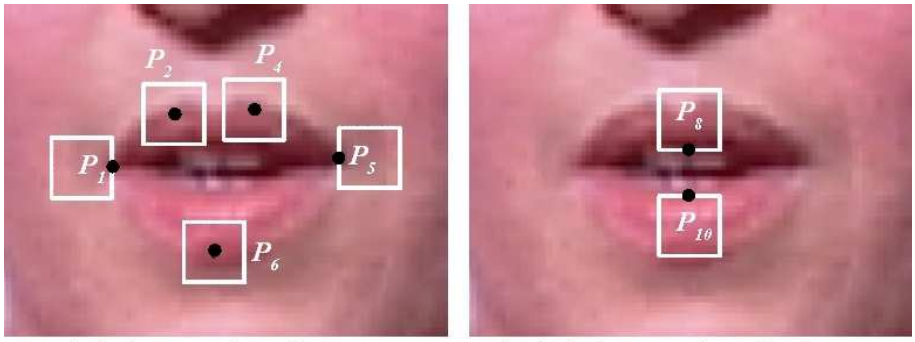
b. Exemple de changement d'apparence au voisinage de la commissure interne gauche

Fig. 6.04. Voisinage des commissures internes au cours d'une séquence.

Pour les 7 autres points clefs, il faut régler les trois paramètres de la méthode de Lucas-Kanade :

1) la position de la fenêtre *W* du voisinage du point à suivre,
2) la taille *m*x*m* de la fenêtre *W*,
3) la condition de fin du processus d'itération.

1) Généralement, le point que l'on souhaite suivre, est placé au centre de la fenêtre *W*. Dans le contexte de la segmentation des lèvres, ce choix n'est pas le plus judicieux dans la mesure où il faut que la région *W* ne change pas trop d'aspect d'une image à la suivante, pour que l'algorithme de Lucas-Kanade soit efficace.
   Considérant les points clefs externes P_2, P_4 et P_6, le voisinage de ces points reste suffisamment constant pour pouvoir les placer au centre des fenêtres *W*. Les changements d'aspect ne concernent principalement que la différence d'illumination d'une image à l'autre, particulièrement pour le point bas P_6, où la région située en dessous de la bouche peut être plus ou moins sombre suivant l'ouverture de la bouche. Dans ce cas, le suivi peut être affecté et il est nécessaire d'effectuer un ajustement (cf. Section 6.1.3.b). En choisissant une taille de fenêtre permettant de ne pas chevaucher l'intérieur de la bouche ($m/2$ < épaisseur de la lèvre), la position des fenêtres de référence *W* est choisie de telle façon que les points suivis soient situés en leur centre (cf. Fig. 6.05.a).
   Concernant les commissures externes P_1, P_5 et les points clefs internes P_8 et P_{10}, le problème est différent dans la mesure où ils se trouvent proches de l'intérieur de la bouche (lorsque la bouche est grande ouverte pour les commissures). Ainsi, le voisinage de ces points peut être affecté par l'apparition des dents, de la langue, des gencives ou de la cavité orale et changer d'aspect d'une image à l'autre. Pour rendre le suivi plus robuste, les points suivis sont placés au milieu d'un des côtés de la fenêtre *W* (cf. Fig. 6.05b). De cette façon, les fenêtres *W* ne chevauchent pas l'intérieur de la bouche.



a. Choix de W pour les points externes b. Choix de W pour les points internes

Fig. 6.05. Positions des fenêtres de référence W utilisées pour suivre les points clefs.

2) La taille mxm des fenêtres W est un compromis entre la complexité du calcul et la précision du suivi. Plus la taille augmente, plus lent est le suivi. Cependant, une valeur trop faible de m implique de considérer un voisinage restreint du point suivi et cela entraîne des appariements plus aléatoires. Pour déterminer m, nous avons utilisé la remarque faite dans 1) : il faut que les fenêtres W des points externes P_2, P_4 et P_6, ne chevauchent pas l'intérieur de la bouche. Ainsi, on prend m/2 plus petit que l'épaisseur moyenne de la lèvre supérieure (généralement moins épaisse que la lèvre inférieure). Par exemple, pour nos images de la base TELMA, la hauteur moyenne de bouche (distance entre P_3 et P_6) est 20 pixels et on a fixé m=12 (l'épaisseur de la lèvre supérieure fait en moyenne 8 pixels au niveau des points P_2 et P_4, là où l'épaisseur est la plus grande). Sur la figure 6.05.b, on peut remarquer qu'avec m=12, les fenêtres W des points clefs internes P_8 et P_{10} peuvent dépasser les frontières extérieures de la bouche. Ceci est particulièrement vrai pour le point haut P_8, car la lèvre supérieure peut avoir une épaisseur plus petite que m. Cela n'affecte pas le suivi dans la mesure où la région au dessus de la bouche et l'épaisseur de la lèvre supérieure restent suffisamment constante d'une image à l'autre.

3) Nous avons choisi deux conditions d'arrêt pour la méthode de minimisation. Tout d'abord, une condition sur la valeur du déplacement incrémental est testée. Si la condition est vérifiée la méthode de minimisation s'arrête, sinon le processus continue jusqu'à atteindre un nombre d'itérations maximum autorisées.

   - La valeur du seuil du déplacement incrémental a été fixée à 0,1 pixel. En dessous de cette valeur, l'amplitude du déplacement incrémental devient négligeable par rapport au déplacement total.
   - Expérimentalement, nous avons vu que la convergence s'effectue généralement rapidement et qu'elle dépasse rarement 5 itérations. Nous avons donc fixé le seuil limite à 6 itérations avant d'obliger le processus à s'arrêter.

6.1.3. Ajustement du suivi des points clefs externes et internes

La figure 6.06 montre un exemple du suivi des 7 points clefs sur quelques images d'une séquence de la base TELMA. D'une image à la suivante, l'algorithme de Lucas-Kanade fournit une bonne estimation de la position des points. Cependant l'erreur s'accumule d'image en image et le suivi devient peu fiable après plusieurs images.

La dégradation du suivi concerne, en particulier, les commissures externes et les points P_8, P_{10} et P_6 :

- Pour P_1 et P_5, l'accumulation d'erreur s'explique par le fait que les commissures ne peuvent être vues comme de véritables points, mais plutôt comme une région où les contours extérieurs des lèvres se rejoignent.
- Pour P_8 et P_{10}, l'erreur de suivi arrive souvent lorsque la bouche s'ouvre (cf. Fig. 6.06).
- Pour le point bas P_6, la difficulté vient du fait que ce point se situe sur un contour qui est généralement horizontal. Dans ce cas, la première composante du gradient g (cf. Eq. 6.08) est négligeable et l'estimation du mouvement horizontal est peu fiable. On remarque sur la figure 6.06 que l'erreur du suivi concerne principalement la position horizontale de P_6.

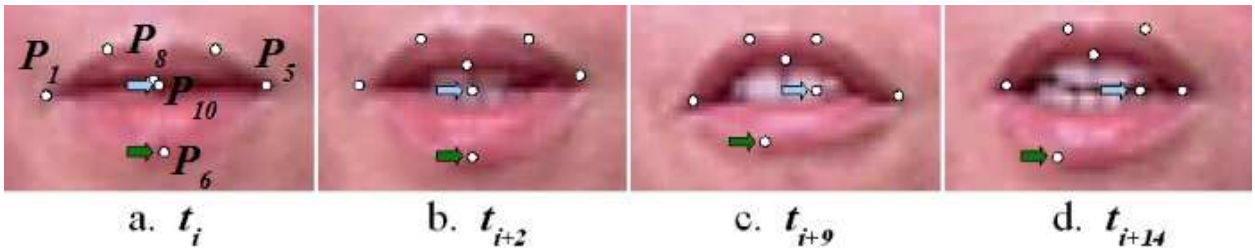


Fig. 6.06. Accumulation des erreurs du suivi des points clefs.

En conclusion, l'algorithme de Lucas-Kanade permet de nous fournir une bonne estimation de la position des points clefs P_1, P_2, P_4, P_5, P_6, P_8 et P_{10} d'une image à la suivante, mais les points ont besoin d'être réajustés à chaque image pour éviter l'accumulation des erreurs de suivi au cours de la séquence.

Pour la suite de l'étude, nous adoptons les notations suivantes :

- $P_i'(t)$ est l'estimation du point P_i obtenue par la méthode de Lucas-Kanade dans l'image courante à l'instant t.
- $P_i(t)$ est la position recalée du point P_i dans l'image courante à l'instant t.

Du fait des propriétés spécifiques des régions entourant ces points clefs, nous avons proposé différentes méthodes d'ajustement en fonction du point à recaler.

6.1.3.a. Ajustement des commissures externes P_1 et P_5

Les positions des commissures externes obtenues par la méthode de suivi ($P_1'(t)$ et $P_5'(t)$) peuvent ne plus être sur la ligne des minima de luminance L_{min}, comme cela a été supposé dans ce travail (cf. Section 5.2.2). La première étape à réaliser est de replacer chacune des deux estimations sur le plus proche pixel appartenant à L_{min}.

Ensuite, il est possible de calculer un modèle déformé du contour extérieur, à partir de la segmentation réalisée sur l'image précédente (à l'instant $t-1$) et des estimations des points clefs externes ($P_1'(t)$, $P_2'(t)$, $P_4'(t)$, $P_5'(t)$ et $P_6'(t)$).

Modèle extérieur déformé :

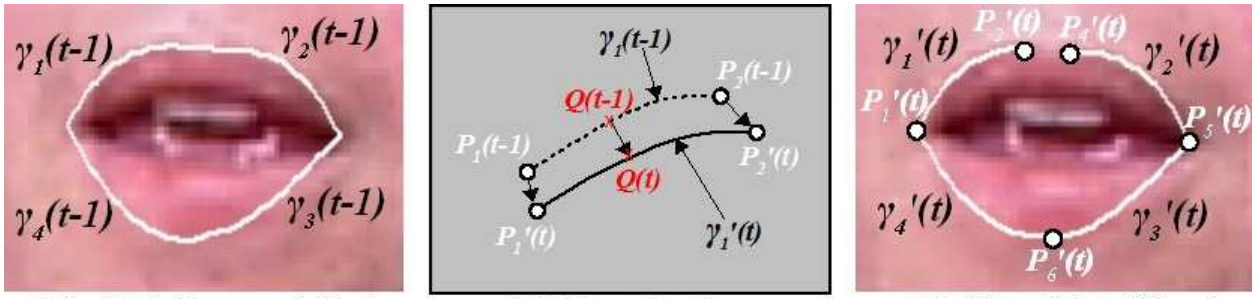
Nous avons à disposition le résultat de la segmentation du contour extérieur des lèvres dans l'image précédente avec les quatre cubiques notées $\gamma_1(t-1)$, $\gamma_2(t-1)$, $\gamma_3(t-1)$ et $\gamma_4(t-1)$ (cf. Fig. 6.07.a). Les cubiques $\gamma_{i=1\ à\ 4}(t-1)$ sont déformées pour coïncider avec les estimations des points clefs externes ($P_1'(t)$, $P_2'(t)$, $P_4'(t)$, $P_5'(t)$ et $P_6'(t)$) et pour obtenir des estimations des cubiques, notées $\gamma_{i=1\ à\ 4}'(t)$, dans l'image courante à l'instant t (cf. Fig. 6.07.c). Chaque point de la cubique à l'instant $t-1$ est déplacé à l'instant t, en utilisant une moyenne pondérée des déplacements des deux points extrêmes de la cubique.
La méthode de déformation du modèle est illustrée avec la cubique $\gamma_1(t-1)$ et les points extrêmes $P_1(t-1)$ et $P_2(t-1)$ (cf. Fig. 6.07.b). Si on note $Q(t-1)$ un point de la cubique $\gamma_1(t-1)$ et $Q(t)$ ce même point déplacé à l'instant t, le déplacement du point Q est obtenu de la manière suivante :

$$d_Q = d_{P_1}\left(1 - \frac{|P_1(t-1)\,Q(t-1)|}{|P_1(t-1)\,P_2(t-1)|}\right) + d_{P_2}\left(1 - \frac{|P_2(t-1)\,Q(t-1)|}{|P_1(t-1)\,P_2(t-1)|}\right) \quad (6.12)$$

où :

$$d_Q = \overrightarrow{Q(t-1)Q(t)}$$
$$d_{P_1} = \overrightarrow{P_1(t-1)P_1'(t)}$$
$$d_{P_2} = \overrightarrow{P_2(t-1)P_2'(t)}$$

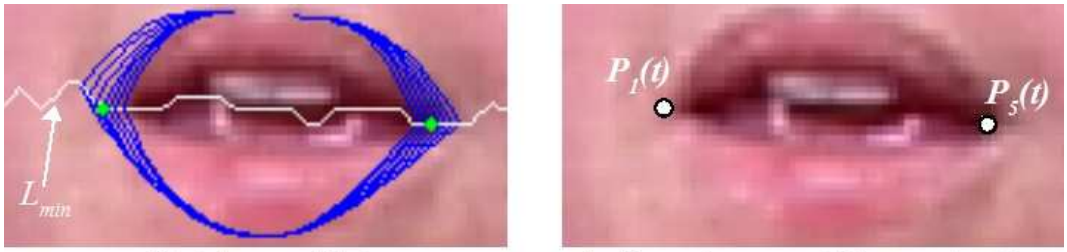
et d_Q, d_{P1} *et* d_{P2} sont les vecteurs déplacements de Q, P_1 et P_2.



a. Résultat de l'image précédente b. Déformation de γ_1 c. Modèle extérieur déformé

Fig. 6.07. Déformation du modèle extérieur de l'image précédente (t-1) à l'image courante (t).

Ce modèle déformé fournit un bon aperçu du contour extérieur à l'instant *t* (cf. Fig. 6.07.c). La méthode utilisée pour ajuster les estimations des commissures externes ($P_1'(t)$ et $P_5'(t)$) consiste à calculer le modèle extérieur déformé pour plusieurs pixels candidats $P_1(t)$ et $P_5(t)$ et à déterminer le meilleur candidat avec la technique de maximisation des flux moyens de gradient, à travers les cubiques déformées obtenues. Nous supposons toujours que les commissures extérieures se trouvent sur L_{min}. Pour chacune des commissures, nous testons sept points appartenant à L_{min} : la position estimée (qui a été initialement replacée sur L_{min}), trois points sur la gauche et trois points sur la droite (cf. Fig. 6.08.a). Pour chaque candidat, nous calculons le modèle déformé. Les positions recalées $P_1(t)$ et $P_5(t)$ sont déterminées respectivement par les deux meilleurs couples de cubiques ($\gamma_1'(t)$, $\gamma_3'(t)$) et ($\gamma_2'(t)$, $\gamma_4'(t)$). Les meilleurs couples de cubiques sont désignés avec la méthode de maximisation des flux moyens de gradient, de la même manière que pour la segmentation statique (cf. Section 5.2.2). Le gradient G_1 est utilisé pour les flux à travers $\gamma_1'(t)$ et à travers $\gamma_2'(t)$, et le gradient G_2 est utilisé pour les flux à travers $\gamma_3'(t)$ et à travers $\gamma_4'(t)$.



a. Test des candidats b. Commissures extérieures ajustées

Fig. 6.08. Ajustement des commissures externes.

6.1.3.b. Ajustement des points clefs externes P_2, P_4 et P_6

L'algorithme de Lucas-Kanade donne les estimations $P_2'(t)$, $P_4'(t)$ et $P_6'(t)$ qui sont soit précises (les points sont sur les contours extérieurs des lèvres), soit proches des contours. Pour recaler ces trois points clefs, nous utilisons les snakes dans leur version standard comme introduits par Kass *et al.* [Kass, 1987] et détaillés dans la section 2.4.1.

Pour rappel, les contours actifs standards sont des courbes qui évoluent, d'une manière itérative, d'une position initiale jusqu'à se coller sur le contour recherché. La convergence du snake se fait en minimisant une fonctionnelle d'énergie composée d'un terme d'énergie externe, lié à l'image (pour attirer la courbe vers les contours), et d'un terme d'énergie interne, qui impose des contraintes de forme de la courbe pendant la déformation).

Comme nous l'avons expliqué lors de la discussion sur les contours actifs standards (cf. Section 3.1), l'étape d'initialisation est capitale et les courbes initiales doivent être proches des contours recherchés pour obtenir un bon résultat de segmentation. Or, nous venons de voir que le modèle déformé, présenté dans la section précédente, est proche des contours extérieurs des lèvres, spécialement avec les commissures externes qui sont désormais recalées. Nous initialisons deux snakes classiques à partir du modèle déformé :

- un snake supérieur est initialisé avec les deux cubiques $\gamma_1'(t)$ et $\gamma_2'(t)$ du modèle déformé (cf. Fig. 6.09.a),
- un snake inférieur est initialisé avec les deux cubiques $\gamma_3'(t)$ et $\gamma_4'(t)$.

Les courbes cubiques sont échantillonnées pour donner les points initiaux composant les deux snakes (cf. Fig. 6.09.b). Pour la convergence des snakes, nous n'utilisons pas d'énergie interne. Les coefficients α et β, qui permettent de paramétrer la définition de l'énergie interne (cf. Eq. 2.07), sont difficiles à évaluer automatiquement et un jeu de paramètres est rarement réutilisable pour d'autres d'images. Dans notre cas, les courbes initiales étant très proches des contours, la convergence des snakes est réalisée en quelques itérations et les courbes n'ont finalement pas besoin d'être régies par des contraintes de forme. Les énergies externes sont basées sur le gradient G_1 pour le snake supérieur et G_2 pour le snake inférieur.

A la fin de la convergence, les points recalés $P_2(t)$, $P_4(t)$ et $P_6(t)$ sont les trois points des snakes finaux supérieur et inférieur les plus proches des estimations $P_2'(t)$, $P_4'(t)$ et $P_6'(t)$ (cf. Fig. 6.09.c). La figure 6.09 illustre l'ajustement des points $P_2'(t)$ et $P_4'(t)$.

a. Modèle extérieur déformé

b. Snake supérieur initial (échantillonnage du modèle déformé)

c. Convergence du snake et recalage des points P_2 et P_4

Fig. 6.09. Ajustement des points clefs externes.

Détection du point clef externe P_3 :

Maintenant que la position des points $P_2(t)$ et $P_4(t)$ est ajustée, nous pouvons détecter la position de $P_3(t)$. Pour cela, on suppose que $P_3(t)$ se trouve sur la colonne médiatrice du segment $[P_2(t)P_4(t)]$. Cinq pixels au dessus et cinq pixels en dessous de l'ordonnée moyenne y_{24} des points $P_2(t)$ et $P_4(t)$ sont testés (cf. Eq. 6.13). Le meilleur candidat maximise le flux moyen du gradient G_1 à travers la ligne brisée $[P_2(t) P_3(t) P_4(t)]$. La figure 6.10 illustre la recherche du point $P_3(t)$.

$$y_{24} = \left(\frac{y_2 + y_4}{2}\right) \tag{6.13}$$

où y_2 est l'ordonnée de $P_2(t)$ et y_4 est l'ordonnée de $P_4(t)$.

a. Image b. Test des candidats $P_3(t)$ c. Points $P_2(t)$, $P_3(t)$ et $P_4(t)$

Fig. 6.10. Détection du point $P_3(t)$.

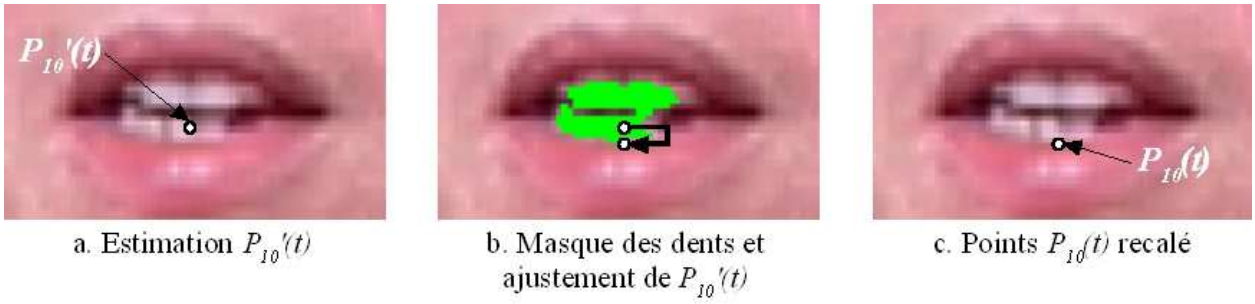
6.1.3.c. Ajustement des points clefs internes P_8 et P_{10}

Les deux points clefs internes P_8 et P_{10} sont des points difficiles à suivre précisément car ils sont sur les frontières de l'intérieur de la bouche. De ce fait, l'environnement de ces deux points change fréquemment d'une image à l'autre, car la bouche alterne continuellement entre l'état ouvert et fermé, et l'apparence de l'intérieur de la bouche varie non-linéairement pendant une conversation (apparition et disparition continues des dents, langue, gencives ou cavité orale). Pour leur ajustement, nous utilisons deux étapes consécutives :

1) ajustement par rapport au masque des dents,
2) ajustement en fonction de l'épaisseur des lèvres.

1) Dans la section 5.3.2.a, nous avons proposé une technique de segmentation des dents. Une fois le contour extérieur connu, l'équation 5.03 permet de déterminer, pour chaque pixel de la bouche, si le pixel est un pixel « dent ». A ce niveau de la segmentation, nous n'avons pas encore le résultat final du contour extérieur de la bouche, mais, une nouvelle fois, nous pouvons utiliser le modèle extérieur déformé présenté dans la section 6.1.3.a.

A partir du masque des dents, l'estimation $P_8'(t)$ est déplacée vers le haut, s'il y a des pixels « dent » au dessus et dans la même colonne que $P_8'(t)$. Et l'estimation $P_{10}'(t)$ est déplacée vers le bas, s'il y a des pixels « dent » au dessous et dans la même colonne que $P_{10}'(t)$. Les points étant sur le contour intérieur, il ne doit y avoir que des pixels « lèvre » au dessus (resp. en dessous) de P_8 (resp. P_{10}). Le processus d'ajustement en fonction des pixels « dent » est illustré pour le point P_{10} sur la figure 6.11.



a. Estimation $P_{10}'(t)$ b. Masque des dents et ajustement de $P_{10}'(t)$ c. Points $P_{10}(t)$ recalé

Fig. 6.11. Ajustement des points clefs internes à partir du masque des dents.

2) A partir des résultats de la segmentation des contours des lèvres obtenus sur les images précédentes de la séquence, nous connaissons l'épaisseur moyenne des lèvres supérieure et inférieure. En supposant que ces épaisseurs ne varient pas trop brutalement d'une image à l'autre (l'épaisseur des lèvres diminue lorsque la bouche s'étire), il est possible d'ajuster les estimations $P_8'(t)$ et $P_{10}'(t)$ en fonction de leur valeur obtenue avec les images précédentes. L'épaisseur de la lèvre la plus haute, notée T_{haut}, correspond à la distance entre les points P_3 et P_8, et l'épaisseur de la lèvre la plus basse, notée T_{bas}, correspond à la distance entre les points P_{10} et P_6 (cf. Fig. 6.12).

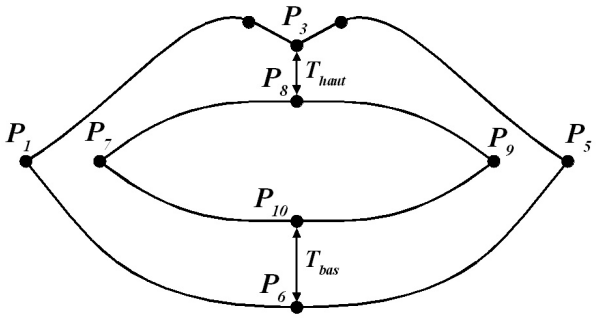


Fig. 6.12. Calcul de l'épaisseur des lèvres.

On calcule l'épaisseur moyenne de la lèvre supérieure, notée $T_{haut}(t_p)$, et l'épaisseur moyenne de la lèvre inférieure, notée $T_{bas}(t_p)$, à partir des cinq images précédentes. Si les valeurs des épaisseurs des lèvres de l'image courante $T_{haut}(t)$ et $T_{bas}(t)$ obtenues à l'aide des points $P_3(t)$ et $P_6(t)$, et des estimations $P_8'(t)$ et $P_{10}'(t)$, ne respectent pas les conditions de l'équation 6.14, les ordonnées des poins clefs internes sont ajustées de la manière suivante :

$$\begin{aligned} T_{haut}(t) < 0.75 * T_{haut}(t_p) \quad \text{or} \quad 1.25 * T_{haut}(t_p) < T_{up}(t) \\ T_{bas}(t) < 0.75 * T_{bas}(t_p) \quad \text{or} \quad 1.25 * T_{bas}(t_p) < T_{low}(t) \end{aligned} \tag{6.14}$$

- l'ordonnée de $P_8'(t)$ est égale à l'ordonnée de $P_3(t)$ plus la valeur de l'épaisseur moyenne $T_{haut}(t_p)$,
- l'ordonnée de $P_{10}'(t)$ est égale à l'ordonnée de $P_6(t)$ moins la valeur de l'épaisseur moyenne $T_{bas}(t_p)$.

Cette méthode de recalage est utile notamment lorsque la bouche s'ouvre trop vite et que l'algorithme de Lucas-Kanade n'a pas réussi à suivre les points clefs internes.

6.2. Test pour la réinitialisation du suivi (Phase 2T)

Lors du suivi des contours des lèvres dans une séquence d'images, il peut arriver que la segmentation échoue pour plusieurs raisons parmi lesquelles :

- un point des modèles paramétriques extérieur ou intérieur a été mal suivi,
- un contour n'est pas assez marqué pour que sa détection soit précise,
- le mouvement de la bouche a été trop rapide d'une image à l'autre,
- la bouche a été partiellement occultée (par exemple par la main pour le projet TELMA).

Si un des cas précédents arrive, le suivi des contours en est affecté. Afin d'éviter que l'algorithme ne diverge, il est nécessaire de réinitialiser l'extraction des contours de la séquence. Le challenge est de savoir quand le suivi a besoin d'être réinitialisé.

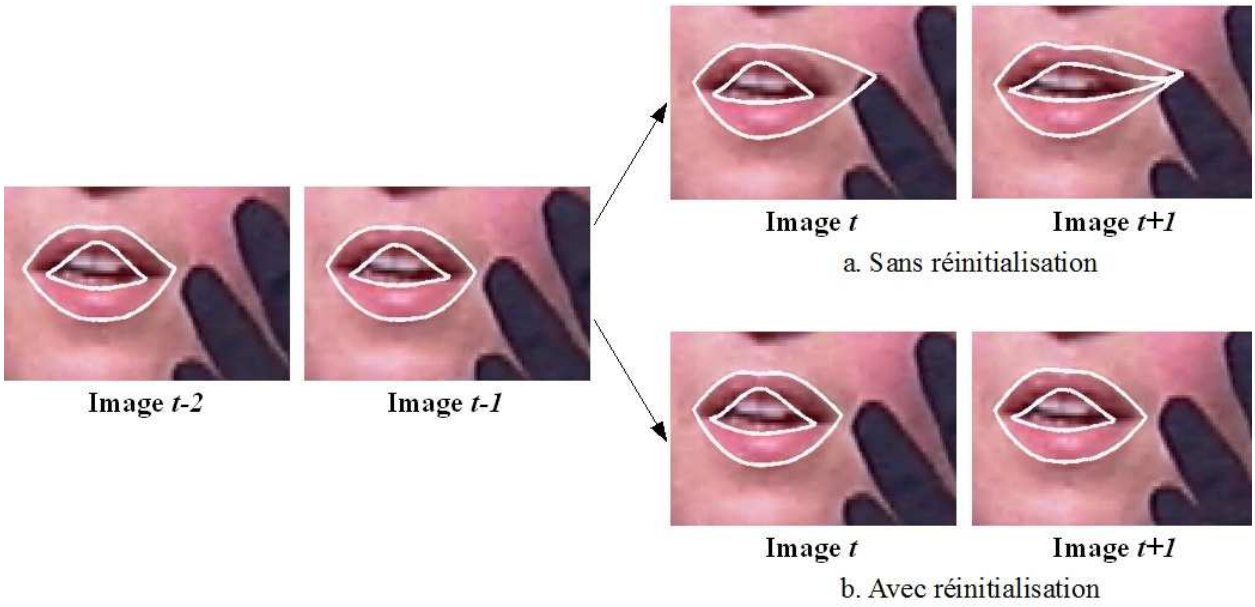


Fig. 6.13. Réinitialisation du suivi au cours de la séquence.

A ce stade de l'algorithme de suivi des contours de la bouche, nous avons la position de 8 points clefs. $P_{11}(t)$ est calculé lorsque la bouche est fermée (cf. Section 6.4) et les commissures internes ($P_7(t)$ et $P_9(t)$) sont détectées en même temps que le calcul des cubiques internes (cf. Section 6.4). En particulier les six points clefs externes $P_{i=1 \text{ à } 6}(t)$ sont connus dans l'image courante; nous nous servons de leur position pour en déduire si le suivi des contours en cours est acceptable ou s'il a besoin d'être réinitialisé. Cette étape correspond à la phase 2T du schéma global de l'algorithme (cf. Fig. 6.01).

Nous comparons les positions des six points clefs externes $P_{i=1 \text{ à } 6}(t)$ de l'image courante avec leur position dans l'image précédente $P_{i=1 \text{ à } 6}(t-1)$. Si la distance entre un des points courants et sa position précédente dépasse un certain seuil, nous en déduisons que la segmentation doit être réinitialisée. Le seuil est fixé expérimentalement, et il prend en compte la taille de la bouche et la cadence d'acquisition des images. Par exemple, pour le projet TELMA, où les bouches ont en moyenne une taille de 50x20 pixels et les images ont été acquises à 50 images/seconde, nous avons fixé le seuil à 10 pixels.

Si la condition n'est pas enfreinte, l'algorithme de suivi continue et il passe aux phases 3T et 4T, qui sont le suivi de la boîte autour de la bouche et l'extraction des contours des lèvres (cf. Section 6.3 et 6.4). En cas de réinitialisation, on applique l'algorithme statique (les phases 1S, 2S et 3S du chapitre 5) sur l'image courante de la séquence qui a posée problème, et cette image est vue comme la première image d'une nouvelle séquence. L'algorithme C3F a besoin du cadre du visage (fournit par CFF) pour déterminer la position des points sur les yeux, le nez et la bouche (cf. Section 5.1.2), mais ensuite, nous gardons le même cadre du visage que celui trouvé lors de la segmentation de la première image de la séquence, afin d'obtenir une certaine continuité pour l'ensemble des résultats du corpus même en cas de réinitialisation.
Sur la figure 6.13, la séquence doit être réinitialisée, car la main est passée près de la bouche ce qui a affecté le suivi de la commissure droite ($P_5(t)$).

L'avantage de la phase 2T est qu'en plus de réinitialiser la segmentation au cours du suivi, elle permet également de compenser une mauvaise segmentation initiale des contours obtenue sur la première image de la séquence. Si l'extraction statique de la première image est inexacte, un des contours au moins n'est pas placé sur une des frontières des lèvres. Le suivi peut continuer, mais le contour mal placé bougera significativement d'une image à la suivante, car il se trouve sur un contour qui n'est pas accentué par les gradients que nous avons développés. Au bout de quelques images, la condition de réinitialisation sur les points clefs externes permet de recommencer l'algorithme de suivi à partir d'une nouvelle image.
La figure 6.14 illustre ce cas particulier. Sur la première image, le point sur le nez, obtenu avec l'algorithme C3F, est trop haut et la boîte autour de la bouche déduite (cf. Chapitre 5) est trop prêt du nez (cf. Fig. 6.14.a). La position du germe extérieur haut, qui initialise le jumping snake extérieur supérieur, ne respecte pas les conditions décrites dans la section 4.2.3.a et le snake ne converge pas sur le contour extérieur des lèvres (cf. Fig. 6.14.b). La segmentation statique donne une mauvaise segmentation pour la première image de la séquence et le suivi est donc mal initialisé. Cependant, une réinitialisation du suivi est faite à l'image 3 et l'algorithme statique fournit cette fois une meilleure segmentation (cf. Bas de la figure 6.14). Nous pouvons remarquer que les commissures intérieures, détectées par l'algorithme statique à l'image 3, sont mal positionnées car les contours sont peu marqués (présence uniquement de la langue et l'intérieur de la bouche a le même aspect que les lèvres). A l'image suivante, le contour est plus marqué et le contour intérieur est bien suivi.

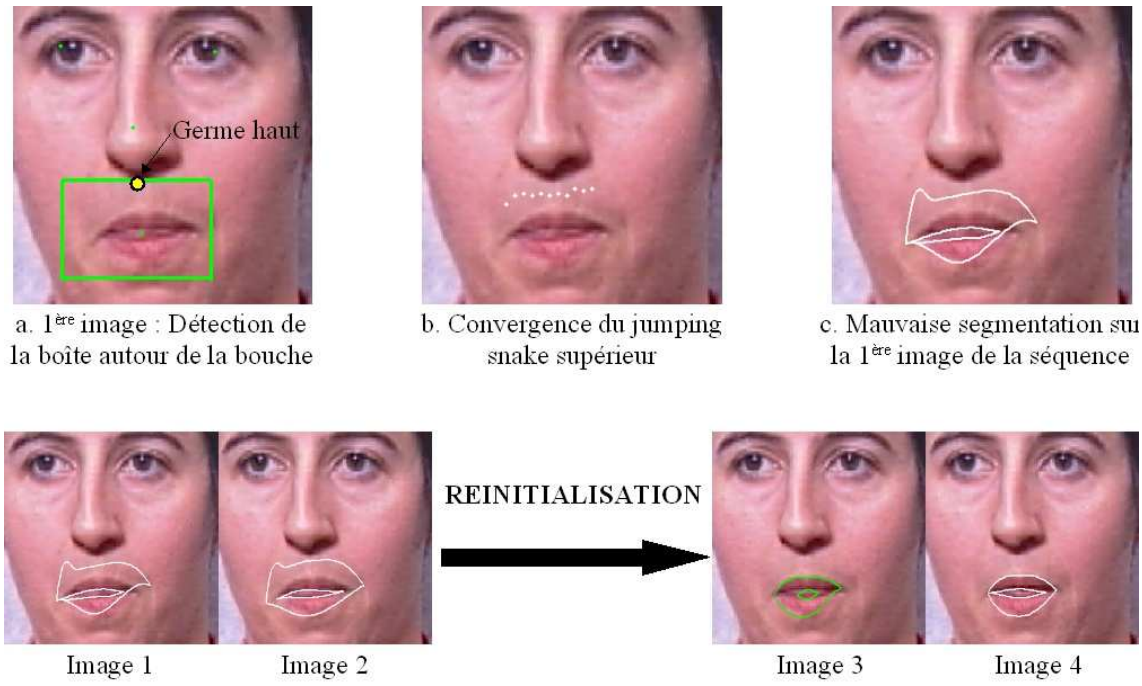


Fig. 6.14. Réinitialisation du suivi lorsque la 1ère image a été mal segmentée.

6.3. Suivi de la boîte autour de la bouche (Phase 3T)

Dans le cas où nous n'avons pas besoin de réinitialiser le suivi des contours, l'algorithme de suivi continue. La phase 3T correspond au suivi de la boîte autour de la bouche, à l'aide d'un filtre de Kalman.

6.3.1. Le filtre de Kalman

Le filtre de Kalman, introduit par Kalman [Kalman, 1960], permet de régulariser la trajectoire d'un modèle. Le filtre de Kalman est un estimateur récursif, c'est-à-dire que l'état courant est défini à partir de l'état précédent et des mesures courantes. L'état précédent est utilisé pour prédire l'état courant et l'observation est utilisée pour affiner la prédiction.

Soient $x \in \Re^n$, le vecteur d'état que l'on cherche à estimer et $z \in \Re^m$, le vecteur des mesures, d'un système qui évolue à temps discret (indice k).
L'état courant x_{k+1} et l'observation z_k sont reliés à l'état du modèle x_k par les relations stochastiques suivantes :

$$x_{k+1} = M_k x_k + B_k u_k + w_k \tag{6.15}$$

$$z_k = H_k x_k + v_k \tag{6.16}$$

où :

- w_k est un bruit gaussien de moyenne nulle et de matrice de covariance Q_k connue; il représente le bruit du modèle
- v_k est un bruit gaussien de moyenne nulle et de matrice de covariance R_k connue; il représente le bruit sur les mesures
- M_k est la matrice d'état qui relie l'état précédent à l'état actuel
- u_k est une entrée de commande connue qui perturbe le système et B_k est la matrice qui relie l'entrée de commande à l'état x
- H_k est la matrice d'observation qui relie l'état actuel à la mesure.

La solution optimale proposée par Kalman [Kalman, 1960] pour estimer l'état courant est fourni par les équations suivantes :

Phase de prédiction :

L'état du système à l'instant k ($x_{k/k-1}$) est prédit avec :

$$x_{k/k-1} = M_k x_{k-1/k-1} + B_k u_k \tag{6.17}$$

et l'estimation de la covariance est :

$$P_{k/k-1} = M_k P_{k-1/k-1} P_k^T + Q_k \tag{6.18}$$

Phase de mise à jour :

Calcul du gain K_k du filtre optimal :

$$K_k = P_{k/k-1} H_k^T \left(H_k P_{k/k-1} H_k^T + R_k \right)^{-1} \tag{6.19}$$

Mise à jour de la matrice de covariance de l'état du système :

$$P_{k/k} = \left(I - K_k H_k \right) P_{k/k-1} \tag{6.20}$$

où *I* est la matrice identité.

Et enfin, l'état estimé est actualisé :

$$x_{k/k} = x_{k/k-1} + K_k \left(z_k - H_k x_{k/k-1} \right) \tag{6.21}$$

6.3.2. Utilisation du filtre de Kalman pour le suivi de la boîte autour de la bouche

Le filtre de Kalman est utilisé lors de la phase 3T de l'algorithme pour suivre la boîte autour de la bouche. Cette section présente les méthodes utilisées pour la phase de prédiction et la phase de mise à jour du filtre. Le filtre de Kalman s'applique aux quatre coordonnées du cadre autour de la bouche et on considère un modèle translationnel; le mouvement des coordonnées d'une image à la suivante est une translation.

Prédiction :

Pour la prédiction de la boîte autour de la bouche, on utilise l'algorithme de block matching. Le bloc de référence est extrait de l'image précédente à partir des coordonnées de la boîte englobante. Il est comparé à plusieurs blocs de même taille dans l'image courante. La recherche est réalisée dans une fenêtre qui correspond à la position du bloc de référence agrandi de ±5 pixels horizontalement et verticalement (cf. Fig. 6.15). Pour comparer la similarité des deux blocs, on utilise l'Erreur Quadratique Moyenne, qui est la différence inter-pixel au carré. La position prédite de la boîte est affectée au bloc donnant l'erreur la plus faible.

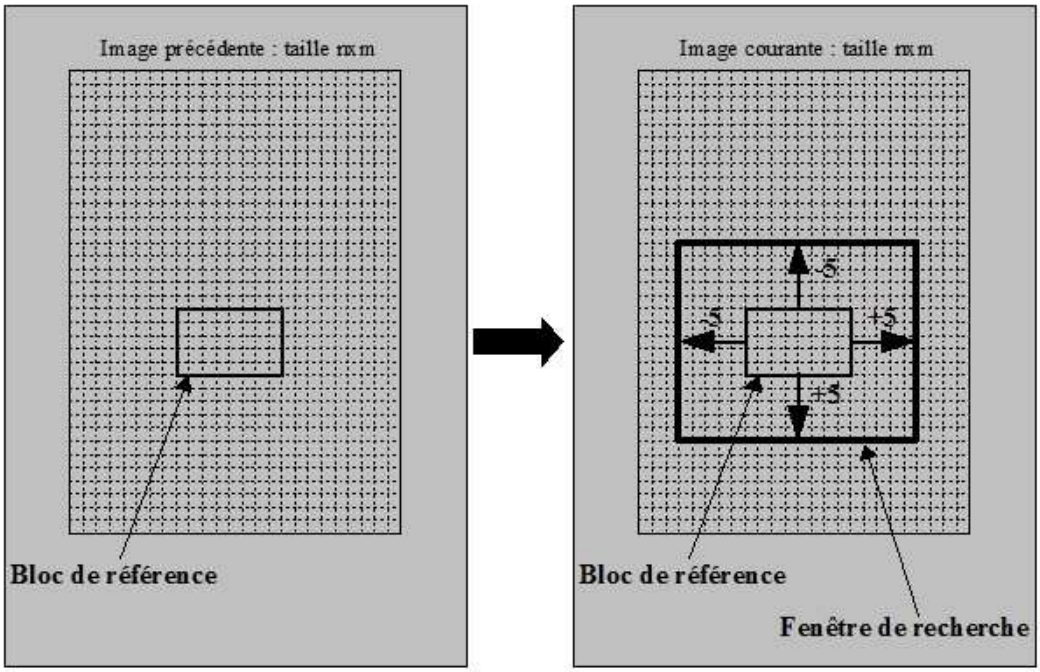


Fig. 6.15. Fenêtre de recherche de l'algorithme block matching.

Mise à jour :

Pour la mise à jour de l'état prédit, le filtre de Kalman utilise une observation. Nous utilisons les positions des points clefs externes $P_{i=1 \text{ à } 6}(t)$ pour calculer une mesure de la boîte courante. Cette boîte observée entoure tous les points clefs d'au moins 5 pixels (cf. Fig. 6.16).
Le filtre de Kalman permet d'obtenir un suivi régularisé de la boîte autour de la bouche pour les images de la séquence. La boîte permet de déterminer la région d'intérêt pour l'extraction des contours. Par exemple, pour le suivi des contours extérieurs, nous calculons plusieurs cubiques et nous déterminons les meilleures pour obtenir le contour final. Si une des cubiques dépasse les limites de la boîte, celle-ci ne sera pas prise en compte (cf. Section 6.4).

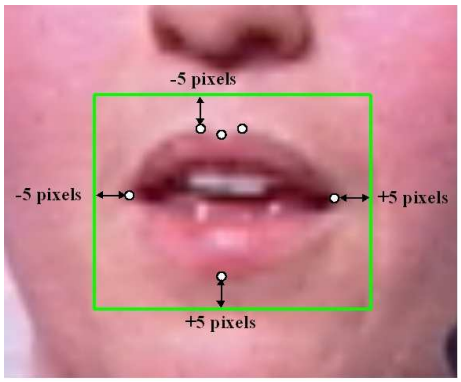


Fig. 6.16. Boîte observée à partir des points clefs externes $P_{i=1 \text{ à } 6}(t)$.

Il est à noter que nous avons essayé d'utiliser la prédiction de la boîte de la bouche pour accélérer la segmentation statique en cas de réinitialisation du suivi (cf. Section 6.2). En effet, nous avons vu précédemment que pour recommencer le suivi des contours après une erreur de segmentation, nous appliquons les 3 phases de l'algorithme statique sur l'image qui a posée problème. Or, la boîte obtenue par le filtre de Kalman permettrait de remplacer les phases 1S et 2S (détection statique du visage et de la boîte autour de la bouche) et de fournir une segmentation statique plus rapide. Malheureusement, il arrive que le suivi doivent être réinitialisé car la bouche s'ouvre trop rapidement et le contour extérieur bas n'a pas suivi. Dans ce cas, le contour extérieur bas se retrouve à l'intérieur de la bouche et la prédiction donne une boîte qui peut aussi se trouver à l'intérieur de la bouche. Comme les germes haut et bas des jumping snakes extérieurs sont déterminés à partir de la boîte et qu'ils doivent se trouver obligatoirement au dessus et en dessous de la bouche, ce cas de figure n'est pas acceptable.

6.4. Extraction des contours des lèvres (Phase 4T)

L'algorithme de suivi utilise les mêmes modèles paramétriques composés de plusieurs courbes cubiques (cf. Section 4.1) pour représenter les contours des lèvres. Les deux modèles extérieur et intérieur sont initialisés à l'aide des points clefs suivis par la méthode de Lucas-Kanade (cf. Section 6.1) .

6.4.1. Extraction du contour extérieur

Nous connaissons la position des six points clefs externes $P_{i=1 \text{ à } 6}(t)$ du modèle paramétrique extérieur. Pour rappel, le modèle est composé de quatre courbes cubiques $\gamma_{i=1 \text{ à } 4}(t)$ et d'une ligne brisée $[P_2(t)\ P_3(t)\ P_4(t)]$ (cf. Fig. 6.17).

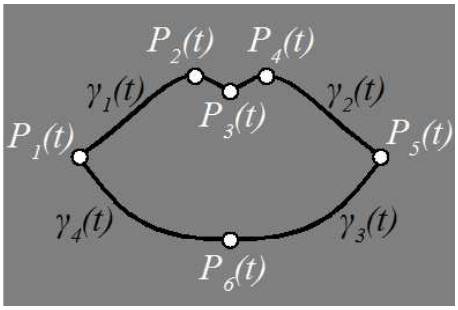


Fig. 6.17. Rappel : modèle paramétrique extérieur.

La ligne brisée est obtenue directement en reliant les trois points $P_{i=2\,à\,4}(t)$, il nous reste donc à déterminer les quatre courbes cubiques.

Une cubique est décrite entièrement si nous connaissons les 4 paramètres qui la régissent. Pour chacune des cubiques du contour extérieur, nous avons la position des deux points extrêmes et le modèle impose une dérivée nulle aux points $P_2(t)$, $P_3(t)$ et $P_4(t)$. En conséquence, nous avons trois équations par cubique et il ne reste plus qu'un seul paramètre à déterminer. Avec l'algorithme de suivi, nous n'avons pas de points supplémentaires sur les contours extérieurs, comme ceux fournis par les deux jumping snakes extérieurs pour la méthode statique. Nous ne pouvons donc pas utiliser d'autres points pour trouver une dernière équation, mais il est possible de se servir des paramètres des courbes cubiques, obtenus lors de la segmentation de l'image précédente, $\gamma_{i=1\,à\,4}(t-1)$.

En supposant que la bouche se déforme suffisamment lentement et que la cadence d'acquisition est suffisamment élevée, nous pouvons considérer que les paramètres des cubiques varient également lentement d'une image à l'autre. En particulier, il est raisonnable de penser que la valeur des pentes des cubiques, au niveau des commissures extérieures P_1 et P_5, sont proches d'une image à la suivante.

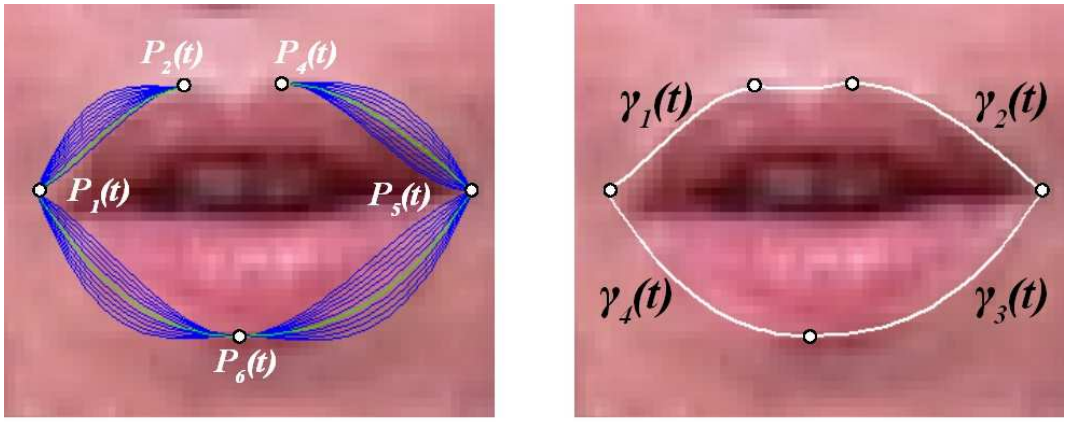
L'optimisation du modèle extérieur et le calcul des cubiques sont réalisés de la manière suivante :

1) Les cubiques sont initialisées avec la position de leurs deux points clefs extrêmes, la contrainte de dérivée nulle au niveau du centre de la bouche et la pente de la cubique de l'image précédente au niveau des commissures extérieures (notée $\rho_i(t-1)$, cf. Fig. 6.18.a).

2) Des cubiques candidates sont testées en faisant varier la valeur des pentes autour de la valeur initiale $\rho_i(t-1)$. Expérimentalement, nous avons établi qu'une dizaine de pentes testées autour de $\rho_i(t-1)$ suffisaient à donner de bons résultats, dans la mesure où les déformations inter-images de la frontière extérieure de la bouche sont petites (cf. Fig. 6.18.b).

3) Les meilleures cubiques sont celles qui maximisent le flux moyen du gradient G_1 (pour $\gamma_1(t)$ et $\gamma_2(t)$) ou G_2 (pour $\gamma_3(t)$ et $\gamma_4(t)$). Les valeurs des pentes de ces cubiques finales $\rho_i(t)$ seront utilisées pour initialiser la recherche dans l'image suivante. Finalement, les quatre meilleures cubiques $\gamma_{i=1\,à\,4}(t)$ et la ligne brisée définissent les contours extérieurs des lèvres (cf. Fig. 6.18.c).

Les différentes cubiques testées doivent se trouver à l'intérieur de la boîte englobante de la bouche pour être prises en compte. Ceci permet de rendre l'algorithme plus robuste vis-à-vis des erreurs de segmentation, lorsque les contours sont peu marqués. Dans le cas d'un contour peu marqué, une cubique candidate $\gamma_i(t)$, ne se trouvant pas sur la frontière des lèvres, pourrait être choisie. Or, dans l'image suivante, la recherche s'effectuera à partir de la valeur de la pente de $\gamma_i(t-1)$, et ainsi de suite, d'image en image. Si la cubique s'éloigne trop de la bouche, aucun des candidats ne se retrouve effectivement sur le contour recherché et la segmentation devient irrécupérable. Le fait de limiter les cubiques à l'intérieur du cadre de la bouche permet de remédier à ce genre d'erreur.



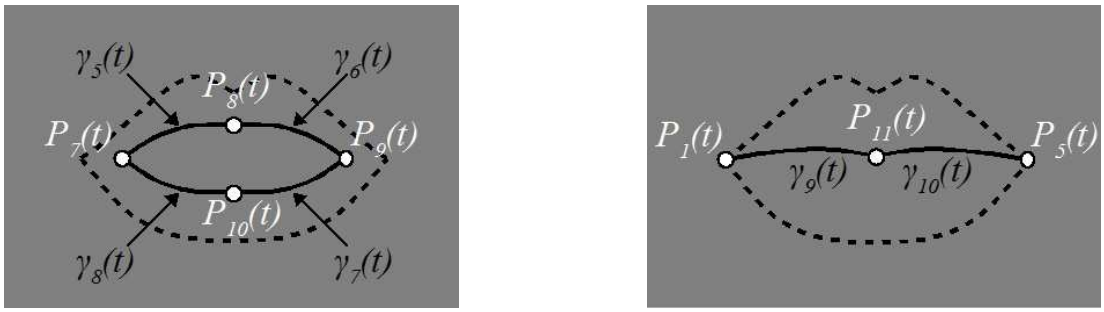
a. Variation des pentes au niveau des commissures extérieures

b. Cubiques extérieures finales

Fig. 6.18. Optimisation du modèle extérieur.

6.4.2. Extraction du contour intérieur

Dans le cas du contour intérieur, l'algorithme de Lucas-Kanade a seulement fourni la position des deux points clefs internes $P_8(t)$ et $P_{10}(t)$. Pour mémoire, le contour intérieur est modélisé par deux modèles paramétriques : un modèle composé de quatre cubiques $\gamma_{i=5\ à\ 8}(t)$ si la bouche est ouverte (cf. Fig. 6.19.a) et un modèle composé de deux cubiques $\gamma_{i=9\ à\ 10}(t)$ si la bouche est fermée (cf. Fig. 6.19.b).



a. Modèle « bouche ouverte »

b. Modèle « bouche fermée »

Fig. 6.19. Rappel : modèles paramétriques intérieurs.

Pour l'extraction du contour intérieur, il faut donc déterminer l'état de la bouche pour choisir le modèle correspondant et déterminer la position des commissures intérieures.

Détection de l'état de la bouche :

De la même manière que pour l'algorithme statique (cf. Section 5.3.1), nous supposons dans un premier temps que la bouche est ouverte et en fonction du résultat de la segmentation du contour intérieur, soit nous validons cette hypothèse, soit nous passons au cas bouche fermée.

Optimisation du modèle intérieur « bouche ouverte » :

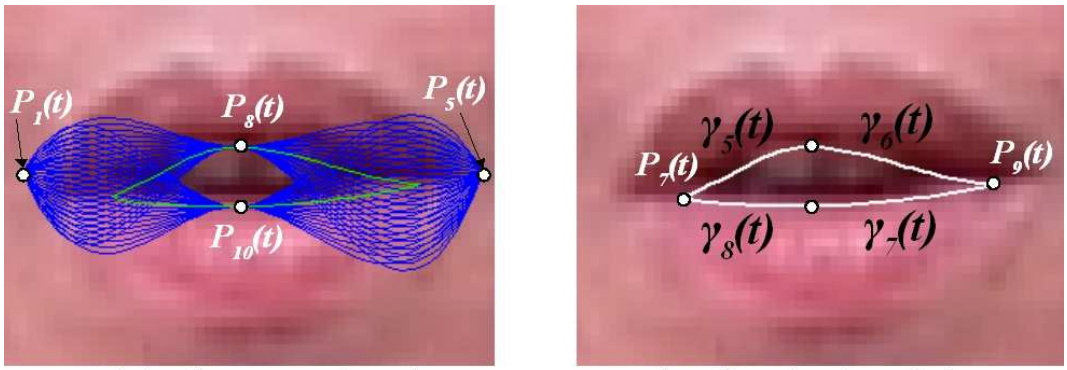
Pour ce modèle, il faut déterminer la position des commissures intérieures $P_7(t)$ et $P_9(t)$. De la même façon que pour la méthode statique (cf. Section 5.3.2), les commissures sont trouvées en même temps que les courbes cubiques. Aussi, comme nous l'avons vu précédemment avec l'optimisation du modèle extérieur, nous utilisons les pentes des cubiques trouvées dans l'image précédente.

L'optimisation du modèle intérieur « bouche ouverte » et le calcul des cubiques sont réalisés de la manière suivante :

1) On suppose tout d'abord que les commissures intérieures sont égales aux commissures extérieures ($P_7(t) = P_1(t)$, et $P_9(t) = P_5(t)$). Les cubiques sont initialisées avec la position de leurs deux points clefs extrêmes (on utilise donc les commissures extérieures), la contrainte de dérivée nulle au niveau du centre de la bouche et la pente précédente au niveau des commissures extérieures ($\rho_i(t-1)$, cf. Fig. 6.20.a).

2) Des cubiques candidates sont testées en faisant varier la valeur des pentes autour de la valeur initiale $\rho_i(t-1)$. Pour le cas intérieur, nous testons une vingtaine de pentes autour de $\rho_i(t-1)$ (deux fois plus que pour le cas extérieur), les déformations inter-images de la frontière intérieure de la bouche sont plus importantes que pour la frontière extérieure (cf. Fig. 6.20.a).

3) Les meilleures cubiques sont celles qui maximisent le flux moyen du gradient G_3 (pour $\gamma_5(t)$ et $\gamma_6(t)$) ou G_4 (pour $\gamma_7(t)$ et $\gamma_8(t)$). Les valeurs des pentes de ces cubiques finales $\rho_i(t)$ seront utilisées pour initialiser la recherche dans l'image suivante. Finalement, les commissures intérieures sont positionnées aux intersections des couples de cubiques ($\gamma_{i=5}(t)$, $\gamma_{i=8}(t)$) et ($\gamma_{i=6}(t)$, $\gamma_{i=7}(t)$) (cf. Fig. 6.20.b), et les quatre meilleures cubiques définissent les contours extérieurs des lèvres.



a. Variation des pentes au niveau des commissures extérieures

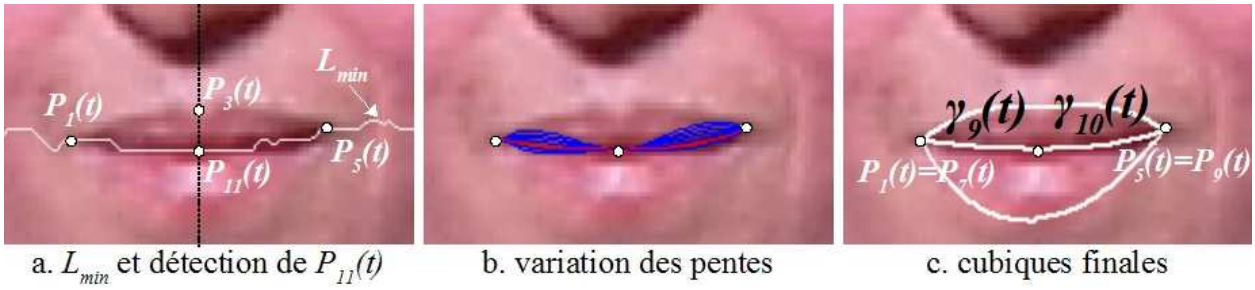
b. Cubiques intérieures finales

Fig. 6.20. Optimisation du modèle intérieur « bouche ouverte ».

Maintenant, il faut vérifier que la bouche était bien ouverte. Le critère de vérification est encore plus simple que pour la méthode statique, dans la mesure où les points clefs recalés $P_8(t)$ et $P_{10}(t)$ fournissent un renseignement fiable sur l'état d'ouverture de la bouche. Si la bouche est fermée, les deux points sont très proches, sinon ils sont d'autant plus éloignés l'un de l'autre, que la bouche est ouverte. Si la surface, définie par les contours intérieurs supérieur et inférieur, est plus faible qu'un certain seuil (fixé expérimentalement à 10 pixels comme précédemment), alors nous en déduisons que la bouche était en réalité fermée et nous passons à l'optimisation du modèle intérieur « bouche fermé ». Sinon la segmentation du contour pour l'image courante est terminée.

Optimisation du modèle intérieur « bouche fermée » :

Si la bouche est détectée fermée dans l'image courante, nous appliquons exactement la même technique que pour une image statique (cf. Section 5.3.3 et Fig. 6.21). Pour rappel, le point clef $P_{11}(t)$ est obtenu à l'aide de la ligne des minima de luminance et il se trouve sur la même colonne que $P_3(t)$. Les commissures internes sont égales aux commissures externes ($P_7(t) = P_1(t)$, et $P_9(t) = P_5(t)$). Deux cubiques initiales $\gamma_9(t)$ et $\gamma_{10}(t)$, sont calculées à l'aide de la méthode des moindres carrés et nous faisons varier la valeur des pentes au niveau des commissures extérieures pour obtenir plusieurs courbes candidates. Les deux meilleures courbes correspondent aux deux maxima des flux moyens du gradient de la luminance. Le modèle paramétrique intérieur est ainsi ajusté et les contours intérieurs sont extraits dans l'image courante.



a. L_{min} et détection de $P_{11}(t)$ b. variation des pentes c. cubiques finales

Fig. 6.21. Optimisation du modèle intérieur « bouche fermée ».

Pour l'image suivante, le cas « bouche ouverte » est obligatoirement traité dans un premier temps. La valeur de la pente au niveau de la commissure extérieure gauche de $\gamma_9(t-1)$ sera utilisée pour, à la fois, initialiser la recherche des cubiques $\gamma_5(t)$ et $\gamma_8(t)$. De même, La valeur de la pente au niveau de la commissure extérieure droite de $\gamma_{10}(t-1)$ sera utilisée pour, à la fois, initialiser la recherche des cubiques $\gamma_6(t)$ et $\gamma_7(t)$.

6.5. Evaluation quantitative des performances de l'algorithme de suivi

6.5.1. Base TELMA

Nous avons vu dans la section 4.4.3 que dans le cadre du projet TELMA, nous avions à disposition un corpus de 250 séquences représentant 82530 images. L'algorithme de suivi a été appliqué automatiquement sur l'ensemble des séquences. Sur les 82530 images, le suivi a été réinitialisé automatiquement 154 fois, essentiellement à cause des mouvements de la main près de la bouche, ce qui correspond à 0,2%, et au moins une réinitialisation a été nécessaire pour 111 séquences sur les 250 disponibles. Ces réinitialisations n'ont pas empêché le suivi de se poursuivre et nous obtenons un résultat pour chacune des séquences.

Plusieurs vidéos résultats sont visibles sur la page : http://www.lis.inpg.fr/pages_perso/stillittano/, dans l'onglet « Résultats et Démo ».

La figure 6.22 montre des exemples de segmentation des contours des lèvres pour plusieurs images issues des séquences du projet TELMA.

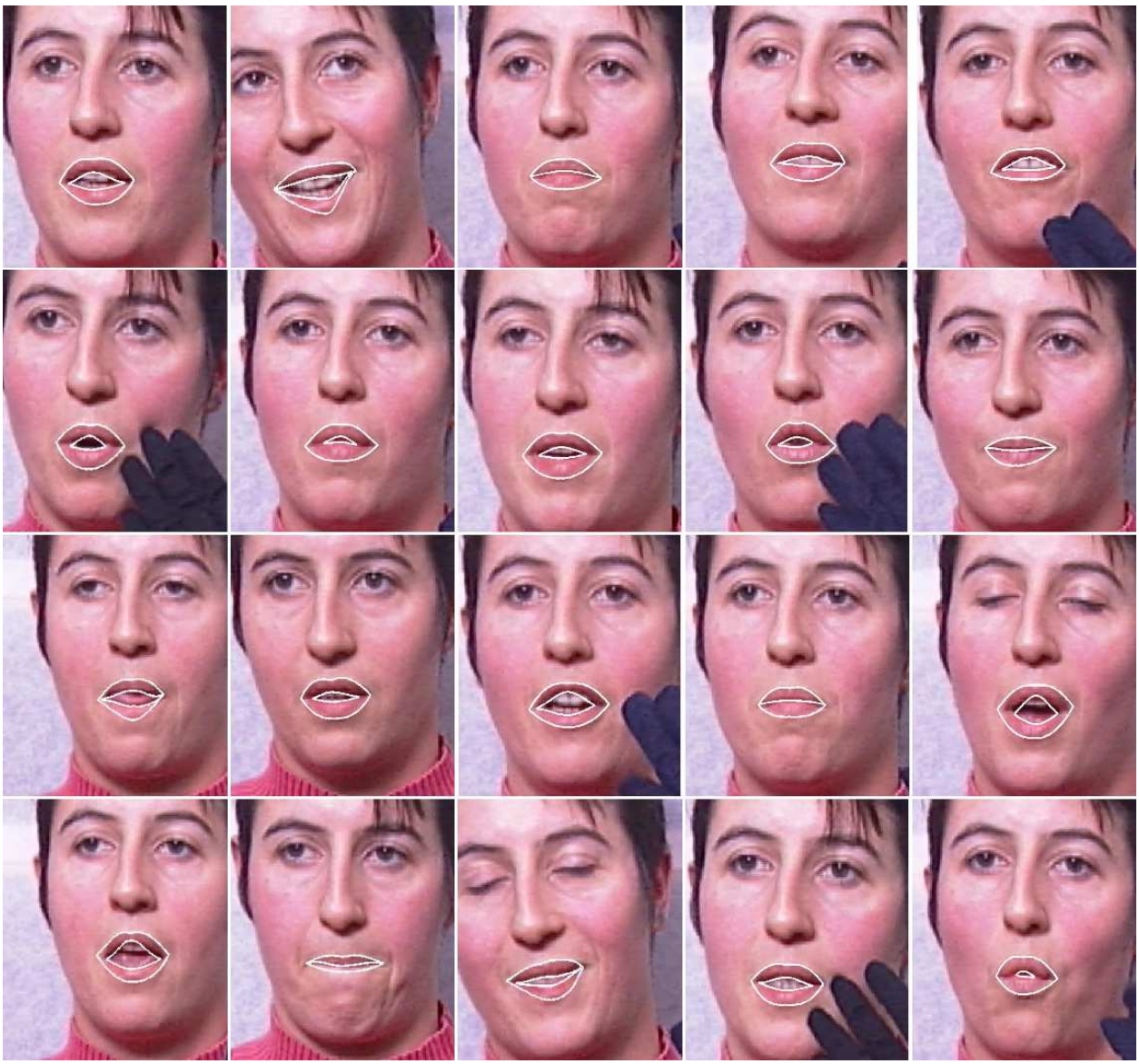


Fig. 6.22. Exemples de résultats du suivi des contours.

6.5.2. Evaluation quantitative

Pour comparer l'expérimentation avec la vérité de terrain, nous utilisons également la méthode introduite par Wu *et al.* [Wu, 2002] (cf. Section 5.4).

6.5.2.a. Constitution de la vérité de terrain

Annoter manuellement les 250 séquences (et donc les 82530 images) aurait été un travail gigantesque et fastidieux. Nous avons décidé de construire une vérité de terrain pour la séquence numéro 69 seulement. Cette séquence a été choisie pour plusieurs raisons :

- c'est la plus longue du corpus : elle est composée de 1412 images
- une réinitialisation est nécessaire au cours du suivi
- la tête de la codeuse se déplace fréquemment au cours de la vidéo, ce qui rend la segmentation plus difficile (cf. Fig. 6.23).

Fig. 6.23. La tête bouge fréquemment sur la séquence 69.

Pour les 1412 images de la séquence numéro 69, nous avons annoté manuellement les contours extérieur et intérieur des lèvres. L'étiquetage des 1412 images a pris en tout environ 47 heures (il faut environ 2 minutes pour annoter une image).

6.5.2.b. Evaluation de la méthode de détection de l'état de la bouche

En plus de l'étiquetage manuel des contours, nous avons constitué une vérité de terrain sur l'état de la bouche pour les 1412 images de la séquence 69. Pour cette séquence, nous avons 732 images où la bouche est ouverte et 680 images où la bouche est fermée.
La méthode de détection de l'état de la bouche proposé (cf. Section 5.3.1) donne le même état que la vérité de terrain pour 1219 images de la séquence, ce qui représente un taux de détections réussies de 86,3%.
Sur les 193 mauvaises détections, 187 concernent l'état fermé et seulement 6 l'état ouvert. Les erreurs de détection plus importantes lorsque la bouche est fermée s'expliquent par le fait que dans ce cas, il arrive fréquemment qu'une zone sombre soit présente entre les lèvres supérieure et inférieure (cf. Dernière image de la figure 6.23). Or, si cette zone sombre est suffisamment grande, l'algorithme de détection fait difficilement la différence avec une bouche légèrement ouverte où l'on voit la cavité orale (cf. Première image de la figure 6.23).

6.5.2.c. Evaluation quantitative de la séquence numéro 69

Les performances de l'algorithme de suivi sont évaluées pour le contour extérieur sur la figure 6.24 et pour le contour intérieur sur la figure 6.25 (lignes noires). Dans les deux cas, nous faisons également une comparaison avec les résultats de segmentation de l'algorithme statique (lignes grises), c'est-à-dire que nous appliquons la méthode statique pour les 1412 images de la séquence et nous comparons les résultats statiques obtenus avec la vérité de terrain. Il est à noter que pour le contour intérieur, nous étudions seulement les 732 images de la séquence où la bouche est ouverte, sont. Pour les 680 images restantes, la bouche est fermée et nous ne pouvons pas appliquer la méthode de comparaison de Wu, car le contour intérieur ne définit pas de zones, mais il est simplement représenté par une ligne.

L'algorithme de suivi (ligne noire de la figure 6.24) fournit de meilleurs résultats que l'algorithme statique (ligne grise de la figure 6.24). En effet, le taux d'erreur de la méthode statique présente plusieurs pics d'erreur qui s'explique par d'importantes erreurs de segmentation pour certaines images. Ces erreurs arrivent lorsque la boîte englobante de la bouche est mal détectée, à cause de la présence de la main près de la bouche. Or les germes des jumping snakes extérieurs sont initialisés à partir de la boîte et si un des germes est trop éloigné des lèvres, alors la segmentation échoue. Le suivi des contours permet une plus grande robustesse (si la première image est bien segmentée ou si la boîte est bien détectée lors d'une réinitialisation). Sinon, lorsque la boîte est bien trouvée, les deux algorithmes fournissent des résultats globalement équivalents. Ce qui est normal, dans la mesure où la même méthode d'optimisation des modèles paramétriques (maximisation de flux moyens de gradients) et les mêmes gradients sont utilisés dans les deux cas.

La figure 6.25 montre de larges erreurs concernant la segmentation du contour intérieur. Ces erreurs arrivent soit lorsque la bouche est considérée comme étant ouverte pour la vérité de terrain alors qu'elle est détectée fermée avec l'algorithme (comparaison entre une zone de plusieurs pixels et une ligne), soit lorsque la bouche est peu ouverte. Aussi, le taux d'erreur pour le contour intérieur est plus important que pour le contour extérieur. Le taux d'erreur moyen pour le contour intérieur est de 0,57 pour la séquence 69, alors qu'il est de 0,23 pour le contour extérieur. Ceci s'explique simplement par le fait que les tailles des zones définies par les contours extérieurs sont beaucoup plus grandes que celles définies par les contours intérieurs. Sur la séquence 69, la taille moyenne (largeur x hauteur) du contour extérieur est de 49 x 25 pixels, alors qu'elle n'est que de 33 x 11 pixels pour le contour intérieur. Avec de si petites aires, même une différence de quelques pixels entre la vérité de terrain et les résultats de l'algorithme donne de larges erreurs.

Par exemple, prenons deux rectangles de taille 33 x 11 et 49 x 25. Supposons que la différence entre les résultats de l'algorithme et de la vérité de terrain n'est que d'un seul pixel sur tout le contour, ce qui correspond au périmètre des rectangles. Dans le premier cas, on a (2x33)+(2x11) = 88 pixels erreurs pour une aire contenant 33x11 = 363 pixels, et le taux d'erreur est 88/363 = 0,24. Dans le deuxième cas, on a (2x49)+(2x25) = 148 pixels erreurs pour une aire contenant 49x25 = 1225 pixels, et le taux d'erreur est 148/1225 = 0,12. Donc pour une même erreur (différence de un pixel entre l'algorithme et la vérité de terrain), on a un taux d'erreur deux fois plus grand dans le premier cas. Il faut donc interpréter avec précaution les résultats de cette évaluation.

En conclusion, une évaluation quantitative n'est pas toujours parlante pour démontrer l'efficacité d'une méthode de segmentation. Surtout que pour une évaluation prenant en compte une vérité de terrain, le résultat dépend fortement de l'expert qui a annoté les contours manuellement, en particulier pour des bouches de petite taille. Il serait préférable d'avoir une vérité de terrain qui serait la moyenne de annotations faites par plusieurs experts. En conséquence, une évaluation de l'algorithme proposé au regard de différentes applications est également présentée dans le chapitre 7.

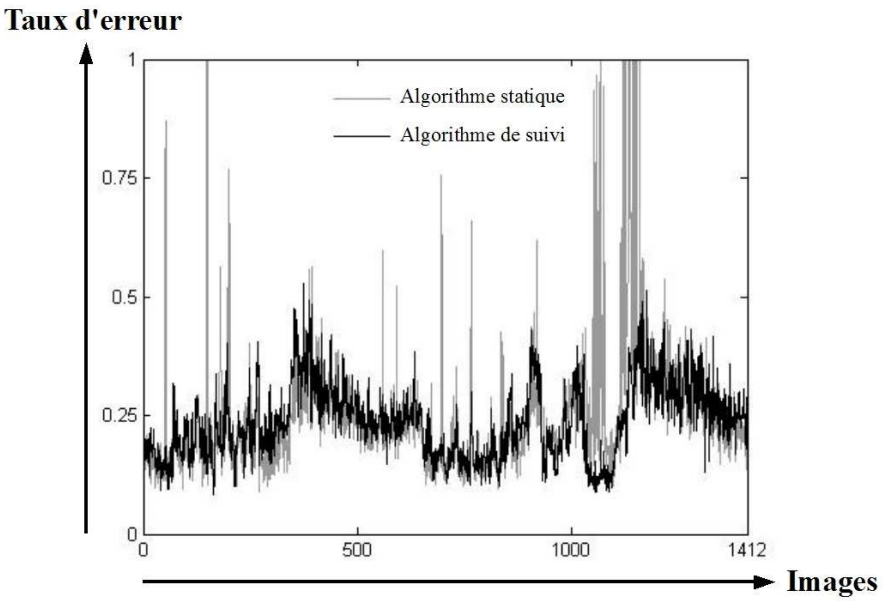


Fig. 6.24. Taux d'erreur pour le contour extérieur des lèvres.

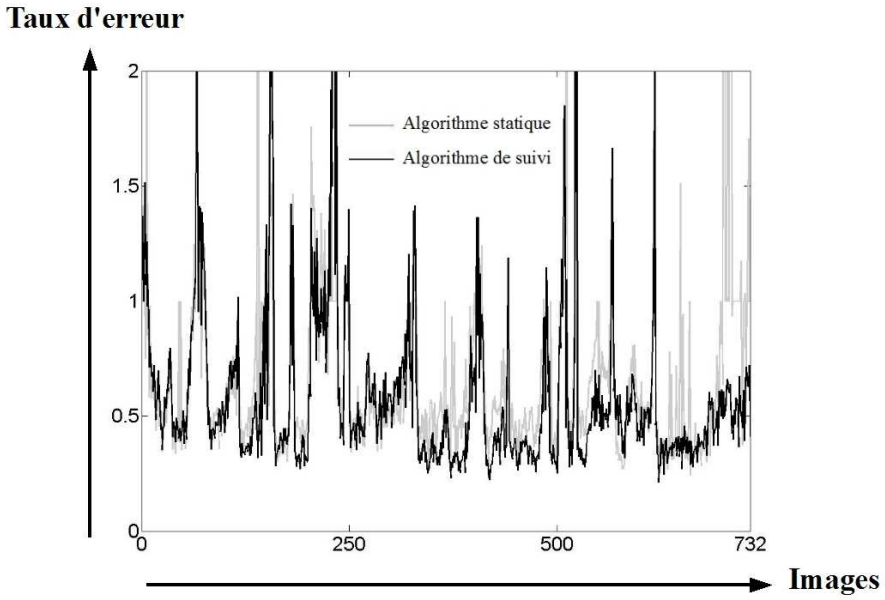


Fig. 6.25. Taux d'erreur pour le contour intérieur des lèvres.

6.5.3. Evaluation de la vitesse de segmentation

La méthode de suivi permet d'augmenter significativement la vitesse de segmentation des contours des lèvres par rapport à la méthode statique. Les tableaux 6.01 et 6.02 détaillent les temps de calcul des différentes étapes pour les deux méthodes proposées. Les temps ont été définis à partir des séquences TELMA (taille 720 x 576 pixels). Le temps d'exécution moyen de l'algorithme statique (cf. Tab. 6.01) est une moyenne des temps calculés pour la première image de chacune des 250 séquences. Les temps d'exécution moyen de l'algorithme de suivi (cf. Tab. 6.02) sont une moyenne des temps calculés sur plusieurs centaines de suivi. Pour les tests, les algorithmes ont été implantés sous MATLAB et ils ont été exécutés avec un microprocesseur Intel Xeon 2.8 GHz.

Phase 1S : Détection du visage	0,27 sec
Phase 2S : Détection de la boîte autour du visage	0,02 sec
Calcul des espaces couleurs	0,19 sec
Calcul des gradients	0,05 sec
Détection du contour extérieur	0,75 sec
Détection du contour intérieur	1,1 sec
Temps d'exécution moyen de l'algorithme statique	2,47 sec

Tab. 6.01. Temps moyen d'exécution pour la méthode statique.

Calcul des espaces couleurs	0,16 sec
Calcul des gradients	0,07 sec
Phase 1T : Suivi des points clefs	0,16 sec
Phase 3T : Suivi de la boîte autour du visage	0,001 sec
Détection du contour extérieur	0,10 sec
Détection du contour intérieur	0,19 sec
Temps d'exécution moyen de l'algorithme de suivi	0,72 sec

Tab. 6.02. Temps moyen d'exécution pour la méthode de suivi.

Nous pouvons constater que l'algorithme de suivi permet une segmentation qui est environ 3,4 fois plus rapide que pour l'algorithme statique. Les modules de détection des contours extérieur et intérieur sont plus rapide en suivi qu'en statique (entre 6 et 7,5 fois plus rapide). Cependant, pour le suivi, ces modules ne consistent qu'à calculer une courbe initiale à partir du suivi des points clefs et à faire varier la pente de la courbe pour sélectionner la meilleure avec le flux moyen de gradient. En statique, ces modules intègrent la convergence des snakes et la recherche des points clefs.
Nous pouvons remarquer que presque 11% du temps de calcul en statique est utilisé pour la détection du visage et 8% pour le calcul des espaces couleurs. En suivi, le calcul des espaces couleurs nécessite 22% du temps de calcul.

6.6. Conclusion

Dans ce chapitre, nous avons proposé un algorithme de suivi des contours extérieur et intérieur des lèvres.

L'algorithme de suivi utilise les mêmes modèles paramétriques que ceux de l'algorithme statique. Un module de suivi des points clefs externes et internes basé sur la méthode de Lucas-Kanade permet d'initialiser rapidement les modèles. Une mesure sur les déplacements effectués par les points clefs externes d'image en image, indique si le suivi doit être réinitialisé, ce qui rend l'algorithme plus robuste vis-à-vis de certains cas difficiles (exemple de la main près de la bouche pour les images de la base TELMA). Un critère géométrique sélectionne automatiquement le modèle approprié pour l'image traitée parmi le modèle développé pour les bouches ouvertes et celui pour les bouches fermées. L'optimisation des courbes du modèle choisi est réalisée en maximisant les flux moyens de gradient.
L'évaluation quantitative de la section 6.5.2 montre que l'utilisation de l'information temporelle permet une augmentation significative de la qualité et de la robustesse de la segmentation par rapport à l'algorithme statique. L'algorithme de suivi propose également une segmentation plus rapide avec un temps de calcul 3,4 fois plus faible que le temps d'exécution de l'algorithme statique.

CHAPITRE 7

Evaluation applicative

A la fin des chapitres 5 et 6, nous avons effectué une évaluation quantitative de l'algorithme statique et de l'algorithme de suivi en comparant les résultats de segmentation avec des vérités de terrain.

Dans ce chapitre, nous proposons de suivre une autre stratégie en réalisant une évaluation en rapport à nos applications cibles : le logiciel Makeuponline et le projet TELMA.

Dans la section 7.1, nous discutons de la segmentation des contours des lèvres dans le cadre du logiciel de maquillage Makeuponline.

Une évaluation applicative du projet de téléphonie à l'usage des malentendants (TELMA) est effectuée dans la section 7.2. L'évaluation consiste à comparer les résultats obtenus par Aboutabit dans ces travaux de thèse [Aboutabit, 2007a], où les paramètres labiaux sont calculés en utilisant des lèvres maquillées en bleu, avec les résultats trouvés par notre algorithme.

Enfin, dans la section 7.3, nous présentons les résultats de travaux sur la parole audiovisuelle en collaboration avec l'Université de Princeton (NJ, USA). L'objectif est d'étudier la corrélation qui existe entre le signal visuel (caractérisé par les contours des lèvres) et le signal audio.

7.1. Evaluation par rapport à l'application Makeuponline

L'algorithme de segmentation des lèvres a pour but d'être intégré au logiciel commercial de mise en beauté Makeuponline développé par Vesalis (cf. Section 1.1).

Actuellement, le produit Makeuponline permet de maquiller virtuellement des images fixes prises par un appareil photo numérique ou une webcam. Pour les lèvres, cela consiste à ajouter du rouge à lèvres et du gloss. Une des conditions demandées est de fermer la bouche, car seul le contour extérieur de la bouche est détecté. Notre algorithme de segmentation doit permettre de lever cette contrainte.
Dans un premier temps, seul l'algorithme statique sera implanté, puis dans une version future, l'algorithme de suivi servira à maquiller les lèvres dans des vidéos.

Malheureusement, l'implémentation de l'algorithme en langage C (nous avons développé notre méthode avec le logiciel Matlab) n'a pas pu être réalisée à temps pour pouvoir tester le rendu de la mise en beauté lorsque la bouche est ouverte en utilisant le produit Makeuponline. Cependant, l'objectif de cette thèse est de pouvoir maquiller les lèvres d'une personne lorsque la bouche est ouverte, sans maquiller l'intérieur de la bouche et notamment les dents. Pour tester cela, nous pouvons maquiller les lèvres à partir de nos résultats sans utiliser directement le logiciel de maquillage virtuel.
La figure 7.01 montre simplement la modification de l'image en appliquant automatiquement une couleur rouge par transparence dans la région segmentée par l'algorithme proposé. Ceci permet d'avoir un aperçu du genre de résultat que nous pourrions obtenir une fois l'implémentation effectuée. Bien entendu, l'ajout de la couleur est ici simpliste, et le maquillage virtuel développé par les graphistes de Vesalis permettra une mise en beauté beaucoup plus réaliste. Toutefois, nous pouvons remarquer que l'objectif est atteint, dans la mesure où à l'aide de l'algorithme de segmentation, il est possible d'ajouter du rouge à lèvre sans maquiller l'intérieur de la bouche. Ainsi, une fois que l'algorithme sera intégré au logiciel Makeuponline, il sera possible de se faire maquiller les lèvres sans obligatoirement avoir la bouche fermée.

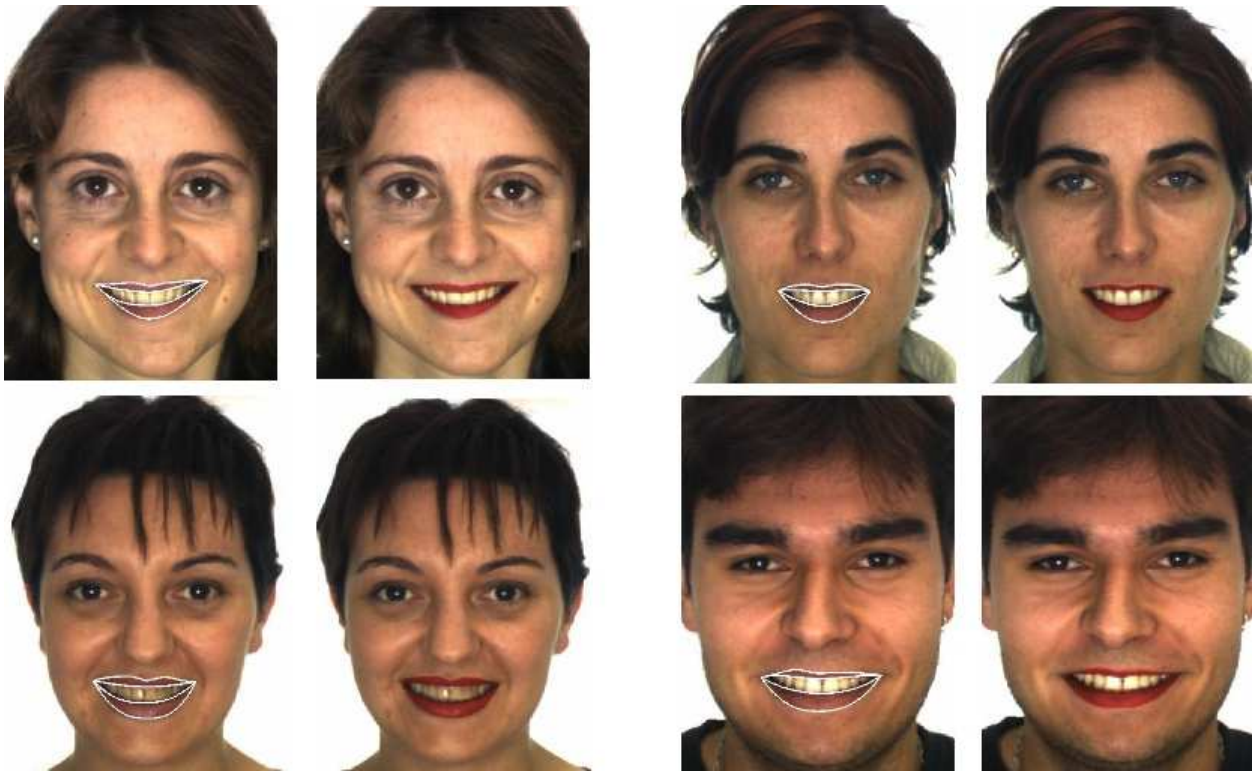


Fig. 7.01. Modification de la couleur des lèvres. Images issues de la base AR [Martinez, 1998].

7.2. Evaluation par rapport au projet TELMA

Nous avons vu dans la section 1.3.2 que les systèmes de reconnaissance automatique de la parole intègrent un traitement visuel permettant d'améliorer le taux de reconnaissance en présence de bruit. L'objectif est d'exploiter les propriétés de la parole audiovisuelle. Le lien entre le canal visuel de la bouche et le canal audio est contenu dans les formes des lèvres. La première étape des applications de lecture labiale consiste à segmenter la région des lèvres et à estimer des paramètres labiaux représentés par différentes mesures calculées à partir des contours extérieur et/ou intérieur. De manière générale, les paramètres labiaux sont (cf. Fig. 7.02, [Aboutabit, 2007b]) :

- A' : ouverture horizontale extérieure de la bouche,
- B' : ouverture verticale extérieure de la bouche,
- S' : aire du contour extérieur de la bouche,
- A : ouverture horizontale intérieure de la bouche,
- B : ouverture verticale intérieure de la bouche,
- S : aire du contour intérieur de la bouche,
- B_{up} : pincement supérieur des lèvres,
- B_{low} : pincement inférieur des lèvres.

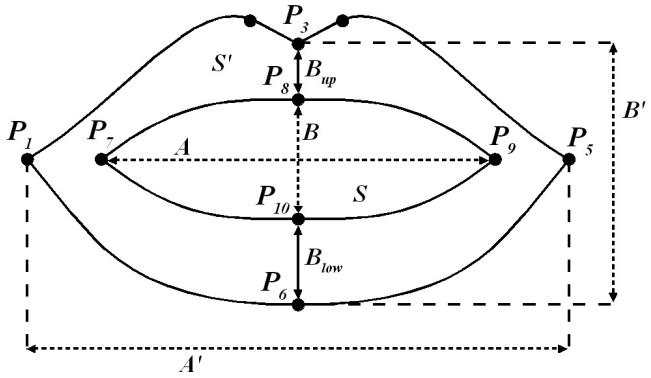


Fig. 7.02. Paramètres labiaux en fonction des points clefs externes et internes des modèles paramétriques.

7.2.1. Evaluation quantitative des paramètres labiaux

Il est possible d'évaluer quantitativement les paramètres labiaux en les comparant à une vérité de terrain. L'évaluation est réalisée à partir de la séquence 69, annotée manuellement (cf. Section 6.5). La vérité de terrain est obtenue en mesurant les distances en fonction des contours annotés manuellement et les paramètres labiaux expérimentaux sont déterminés en fonction des points clefs externes et internes de nos modèles paramétriques des lèvres (cf. Fig. 7.02) de la manière suivante :

- A' est estimé par la distance entre les points clefs externes P_1 et P_5,
- B' est estimé par la distance entre les points clefs externes P_3 et P_6,
- A est estimé par la distance entre les points clefs internes P_7 et P_9,
- B est estimé par la distance entre les points clefs internes P_8 et P_{10},
- B_{up} est estimé par la distance entre les points clefs P_3 et P_8,
- B_{low} est estimé par la distance entre les points clefs P_{10} et P_6.

Les aires S et S' sont obtenues en calculant le nombre de pixels présents à l'intérieur des régions définies respectivement par le contour extérieur et le contour intérieur.

Pour comparer les deux types de données disponibles, nous calculons le pourcentage moyen d'erreur de chaque paramètre labial sur les 1412 images de la séquence. Les résultats de l'évaluation sont répertoriés dans le tableau 7.1.

Labial parameters	***A'***	***B'***	***S'***	***A***	***B***	***S***	B_{up}	B_{low}
Error % (standard deviation)	8 (5)	8 (5)	9 (6)	17 (19)	20 (20)	26 (32)	22 (13)	10 (9)

Tab. 7.01. Évaluation quantitative des paramètres labiaux.

De la même manière que pour l'évaluation quantitative de la section 6.5 réalisée en comparant directement les contours obtenus expérimentalement et la vérité de terrain, nous pouvons observer que l'erreur est plus importante pour les paramètres labiaux internes que pour les paramètres externes. Ceci s'explique également par le fait que les distances étant plus petites pour les valeurs internes, une même erreur (en nombre de pixels) a des répercussions plus grandes pour les paramètres internes que pour les paramètres externes.

7.2.2. Evaluation applicative des paramètres labiaux

La seconde étape des applications de lecture labiale est la reconnaissance des phonèmes en utilisant des paramètres labiaux. Dans le cadre du projet TELMA, nous évaluons notre algorithme de segmentation en comparant nos résultats avec le travail réalisé dans la thèse de Aboutabit [Aboutabit, 2007a] (cf. Section 1.2.3).

Dans [Aboutabit, 2007a], Aboutabit étudie la fusion des informations labiales et gestuelles du code LPC. Une partie de l'étude consiste à faire de la reconnaissance de voyelles à partir des formes labiales. Le flux labial est composé des variations temporelles des paramètres caractéristiques des contours des lèvres. Pour les voyelles, Aboutabit part de l'hypothèse qu'un seul instant de mesure est suffisant pour caractériser une voyelle, puisque toutes les voyelles sont articulées aux lèvres (cf. Fig. 7.02). De plus, dans le cas des voyelles, les paramètres dérivés du contour interne des lèvres ont démontré leur efficacité [Benoit, 1992] [Robert-Ribès, 1995]. La base d'images utilisée dans le travail de Aboutabit est similaire à celle de la base TELMA présentée dans la section 4.4 (même codeuse, même système d'acquisition), excepté le fait que les lèvres sont maquillées en bleu, afin de segmenter facilement la bouche par un seuillage de l'information de chrominance (cf. Fig. 7.03).

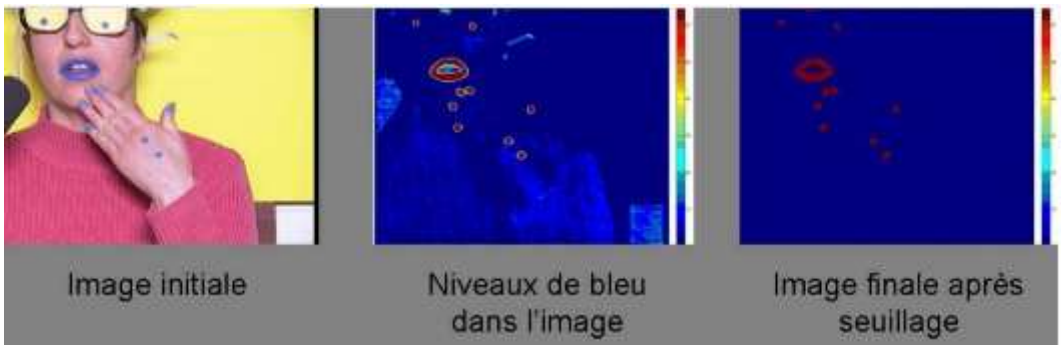


Fig. 7.03. Exemple de détection des éléments bleus de l'image [Aboutabit, 2007a].

Pour analyser la discrimination des voyelles à partir des paramètres labiaux, un arbre hiérarchique de clusters (appelé dendrogramme) est calculé à partir de la distribution (*A*, *B*, *S*) des voyelles à l'instant d'atteinte de la cible labiale. Le dendrogramme consiste en un ensemble de plusieurs axes verticaux connectant des objets (voyelles ou groupes de voyelles) dans un arbre hiérarchique. La hauteur de chaque axe vertical représente la distance entre les deux objets connectés, en utilisant une distance de

Mahalanobis (cf. Fig. 7.04 pour une illustration dans le cas des voyelles).

Distance de Mahalanobis :

Etant donnés deux objets *X* et *Y*, représentés respectivement par deux matrices N_x lignes$*M$ colonnes et *Ny* lignes$*M$ colonnes, la distance de Mahalanobis entre *X* et *Y* est une matrice *D* de N_x lignes$*$ *N*y colonnes. Si nous considérons :

$$D = \{ D(x_i, y_i) \}, \tag{7.01}$$

avec x_i : le *i*éme élément de la matrice *X*, y_j : le *j*éme élément de la matrice *Y*, et *C* : la matrice *M* $*M$ de covariance calculée à partir des données *X* et *Y*, alors la distance de Mahalanobis entre x_i et y_i est définie par :

$$D(x_i, y_i) = \sqrt{(x_i - y_i).C^{-1}(x_i - y_i)^T} \tag{7.02}$$

Si *C* est diagonale, la distance de Mahalanobis correspond à la distance Euclidienne standardisée (c'est-à-dire normalisée par l'écart-type). Si *C* = I, la matrice identité, la distance de Mahalanobis correspond à la distance Euclidienne.

Dans [Aboutabit, 2007a], la distance entre deux objets (groupes de voyelles) correspond à la distance de Mahalanobis la plus petite définie comme :

$$min(D(x_i, y_i)), \;\; i \in [1, N_x], \; j \in [1, N_y] \tag{7.03}$$

Enfin, pour construire l'arbre hiérarchique, cette distance est calculée pour chaque paire d'objets et les distances sont ordonnées par ordre croissant. Les premiers niveaux du regroupement sont donc utilisés pour définir les groupes d'objets (les visèmes des voyelles dans notre cas).

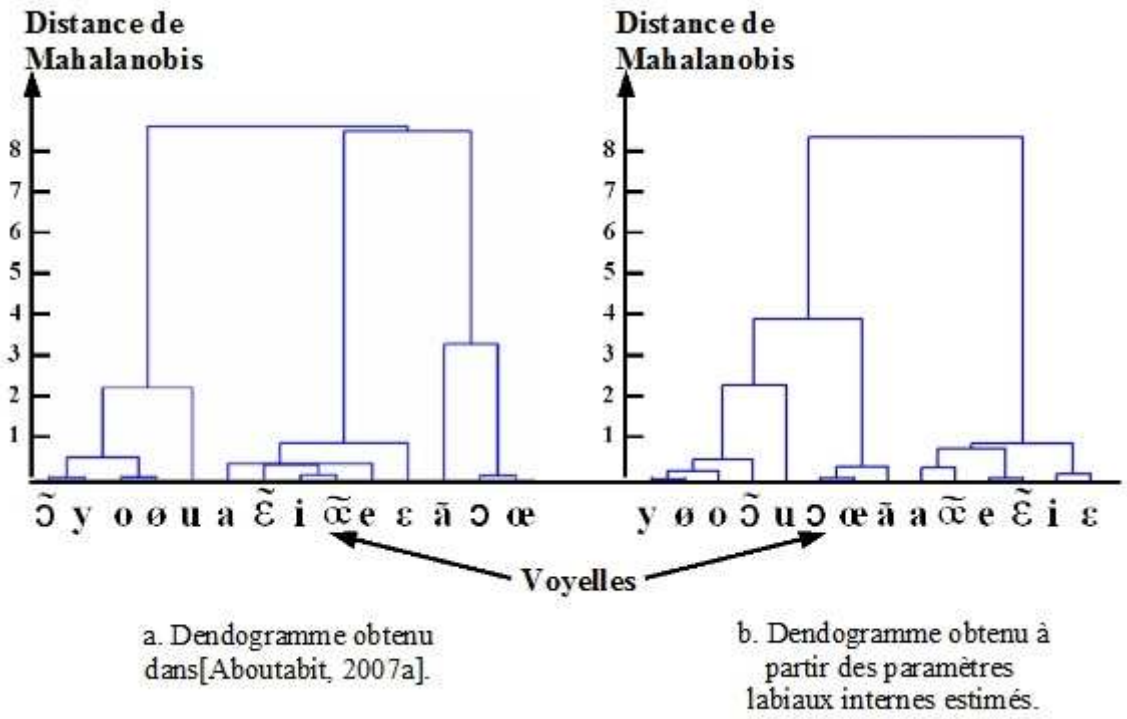
Dendogrammes des groupes de voyelles :

A partir de l'étiquetage phonétique du signal acoustique de l'ensemble des séquences de la base TELMA, les voyelles des phrases sont extraites. Les 14 voyelles de la langue française sont répétées un certains nombres de fois (cf. Tab. 7.02).

a	o	œ	ɛ̃	ø	i	ã	ɔ̃	ɛ	u	ɔ	œ̃	y	e
216	63	24	26	110	176	67	41	96	69	32	26	97	124

Tab. 7.02. Répétitions des voyelles dans le corpus des séquences TELMA de [Aboutabit, 2007a].

Les paramètres labiaux internes (*A*, *B*, *S*) sont extraits aux instants cibles et le dendrogramme des voyelles est construit (cf. Fig. 7.04.a). Nous réalisons la même étude à partir de la base TELMA où les lèvres ne sont pas maquillées. Les paramètres labiaux internes sont obtenus à l'aide de l'algorithme de suivi des contours des lèvres proposé dans cette thèse, et nous construisons également un dendrogramme des voyelles (cf. Fig. 7.04.b). L'idée est de comparer le dendrogramme de Aboutabit (cf. Fig. 7.04.a) obtenue avec des données labiales idéales et celui que nous avons élaboré (cf. Fig. 7.04.b) à partir des données labiales issues de notre processus de segmentation et de suivi des lèvres.



a. Dendogramme obtenu dans[Aboutabit, 2007a].

b. Dendogramme obtenu à partir des paramètres labiaux internes estimés.

Fig. 7.04. Dendrogrammes des voyelles.

Rappelons qu'en ordonnée des dendrogrammes, la longueur des branches verticales est la distance entre deux objets *X* et *Y*. Par exemple, pour le calcul de la distance entre les deux voyelles [ɔ] et [œ] (premier regroupement visible sur la figure 7.04.a), l'objet *X* est l'ensemble des triplets *(A, B, S)* des 32 réalisations labiales de la voyelle [ɔ] et *Y* est l'ensemble des triplets *(A, B, S)* des 24 réalisations labiales de la voyelle [œ] (cf. tab. 7.02). L'ensemble *D* (cf. eq. 7.01) est composé des 32×24 distances deux-à-deux entre les réalisations de [ɔ] et celles de [œ]. La matrice de covariance *C* de dimension 3×3, utilisée dans l'expression de *D* (cf. Eq. 7.02), est calculée sur *(A, B, S)* à partir des 32 réalisations de [ɔ] et des 24 de [œ]. La distance de Mahalanobis, utilisée dans le dendrogramme de la figure 7.04.a, est la plus petite des distances de l'ensemble *D* (cf. Eq. 7.03). Le même calcul est réalisé pour chaque couple de voyelles.
En abscisse des dendrogrammes, nous avons les 14 voyelles de la langue française.

Le dendrogramme de référence (Fig. 7.04.a) montre que les voyelles sont regroupées en trois groupes en accord avec la description phonétique classique des voyelles :

- les voyelles antérieures non arrondies [a, $\tilde{\varepsilon}$, i, $\tilde{\text{œ}}$, e, ε]
- les voyelles arrondies fermées ou moyennement fermées [õ, y, o, ø, u]
- les voyelles arrondies ouvertes [ã, ɔ, œ]

Les mêmes trois groupes sont visibles sur le dendrogrammes construits à partir des paramètres labiaux estimés avec notre algorithme de suivi (cf. Fig. 7.04.b). On peut toutefois remarquer que la distinction entre les groupes « voyelles arrondies ouvertes » et « voyelles arrondies fermées ou moyennement fermées » s'effectue à une distance de Mahalanobis moins importante. Cette différence s'explique par une plus grande difficulté pour notre algorithme à segmenter des bouches peu ouvertes, plutôt que des bouches fermées ou grandes ouvertes.
Les résultats sont donc en accord avec le travail effectué dans [Aboutabit, 2007a] et montrent que notre algorithme peut être utilisé dans des applications de lecture labiale afin d'estimer des paramètres labiaux.

7.3. Travail de collaboration avec l'Université de Princeton

Les travaux présentés dans cette section proviennent d'une collaboration avec l'Institut de Neuroscience et le Département de Psychologie de l'Université de Princeton (NJ, USA). Dans [Chandrasekaran, 2009], Chandrasekaran *et al.* utilisent l'algorithme de suivi proposé dans cette thèse pour caractériser la corrélation qui existe entre le signal vidéo et le signal audio.

Des études dans le domaine des sciences cognitives sur la perception de la parole audiovisuelle ([Sumby, 1954]; [Neely, 1956]) ont montré les bénéfices que pouvait apporter la modalité visuelle sur la perception de la parole. Selon les conditions d'expérimentation, l'information visuelle peut amener une amélioration du rapport signal sur bruit de 15 dB. Si des études à l'échelle des neurones ont permis de localiser les parties du cerveau qui semblent abriter les mécanismes produisant cette amélioration de l'intelligibilité de la parole, les connaissances sur les mécanismes eux-mêmes sont réduites. Ceci est dû principalement au manque d'information sur la structure statistique de la parole audiovisuelle. Des données telles que la corrélation entre les signaux audio et vidéo ou la dynamique temporelle sont limitées.
Dans [Chandrasekaran, 2009], les auteurs étudient la corrélation entre le signal audio et le signal vidéo. Les données utilisées sont le corpus GRID [Cooke, 2006] qui est constitué de 34 locuteurs différents, chacun prononçant 1000 phrases en anglais. Chaque phrase contient six mots (1. une commande, 2. une couleur, 3. une préposition, 4. une lettre, 5. un digit et 6. un adverbe); « bin blue at A1 again » et « place green by D2 now » sont deux exemples possibles. L'audio et la vidéo sont disponibles pour chaque phrase; l'enregistrement vidéo a été effectué à 25 images/seconde.

7.3.1. Analyse du signal visuel

Dans [Chandrasekaran, 2009], les paramètres labiaux ont été obtenus avec l'algorithme de suivi proposé dans cette thèse pour 775 phrases et pour 20 sujets différents. Les paramètres labiaux permettent d'obtenir une estimation de l'aire d'ouverture de la bouche (en pixel carré). La figure 7.05.a montre un exemple de l'analyse de trois images de la séquence « bin red by Y2 now » et la figure 7.05.b illustre l'aire de la bouche en fonction du temps pour la même séquence.

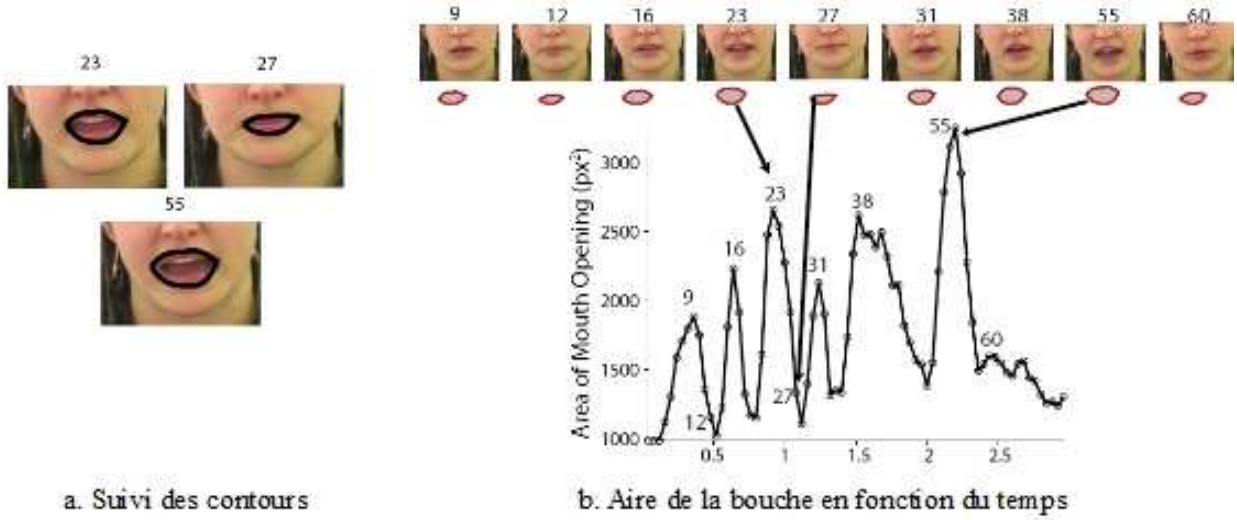


Fig. 7.05. Analyse du signal visuel [Chandrasekaran, 2009].

7.3.2. Analyse du signal audio

La figure 7.06 illustre l'estimation de l'enveloppe large-bande du signal audio. Le signal est tout d'abord filtré par un banc de filtres (dont les fréquences sont espacées de manière logarithmique, cf. Fig.

7.06.b). Puis, la transformée de Hilbert fournit les enveloppes pour chaque bande de fréquence (cf. Fig. 7.06.c). L'enveloppe large-bande du signal audio est obtenue en sommant les enveloppes bande-étroites (cf. Fig. 7.06.d). Le taux d'échantillonnage pour les données issues du corpus GRID est égal à 50 Khz et l'échelle du banc de filtres est 100 Hz – 10 Khz.

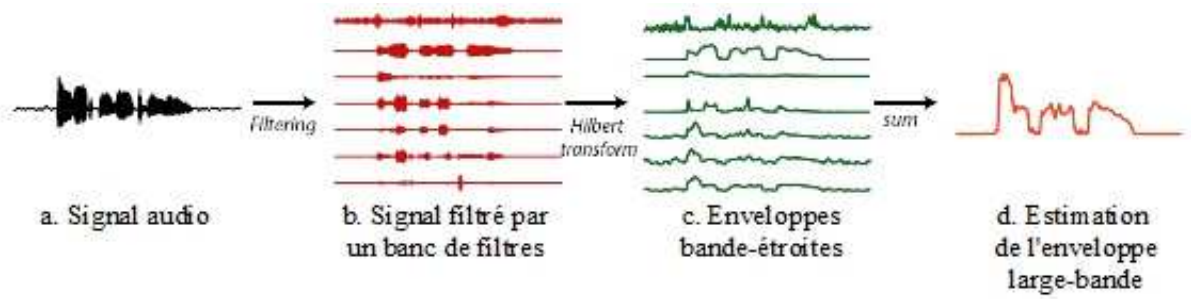


Fig. 7.06. Analyse du signal audio [Chandrasekaran, 2009].

7.3.3. Résultats

7.3.3.a. Corrélation entre le signal visuel et le signal audio

La figure 7.07.a montre l'aire de la bouche et l'enveloppe du signal audio en fonction du temps pour la séquence 1. Les deux signaux sont corrélés (coefficient de corrélation R = 0,742). La même analyse a été effectuée pour l'ensemble des 20 sujets (cf. Fig. 7.07.b) avec deux types de corrélation (intact et shuffled corrélation, voir [Chandrasekaran, 2009] pour plus de détails), qui montre l'existence d'une forte corrélation entre les signaux audio et vidéo.

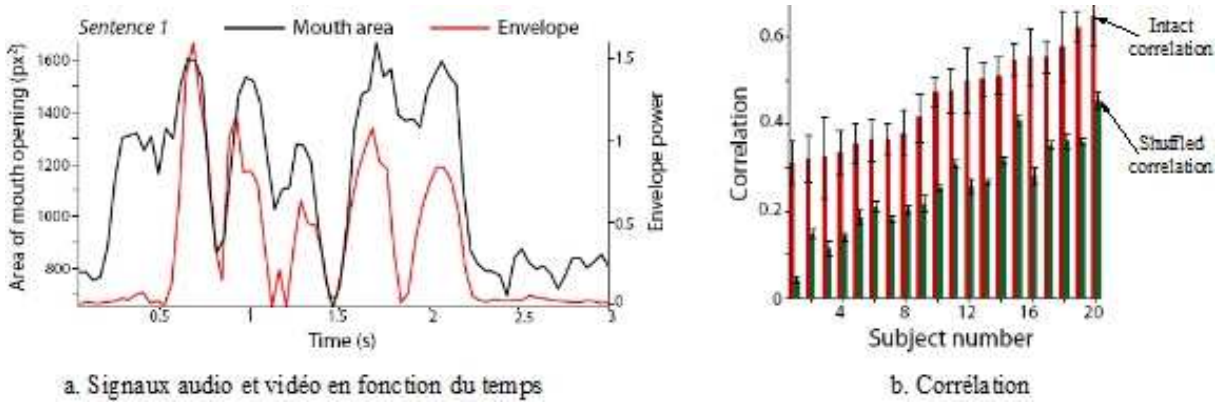


Fig. 7.07. Corrélation entre les signaux audio et visuel [Chandrasekaran, 2009].

7.3.3.b. Relation entre les structures spectrales audio et visuel

Typiquement, la parole est constituée de formants et d'énergie dans plusieurs bandes de fréquence. La figure 7.08 montre la relation qui existe entre les signaux audio et visuel dans les différentes bandes spectrales pour l'ensemble des 20 sujets du corpus GRID étudiés. Cette relation met en évidence l'existence de deux pics pour chaque signal, un au niveau des fréquences 300-600 Hz et un autre au niveau de la fréquence 3KHz. Cette étude suggère que les formants F1 (~300 à 800 Hz) et F2-F3 (~3 Khz) sont corrélés au signal visuel. Les formants étant importants pour caractériser les sons tels que les voyelles, ceci met en évidence l'apport de l'information visuelle pour la reconnaissance de la parole en environnement bruité.

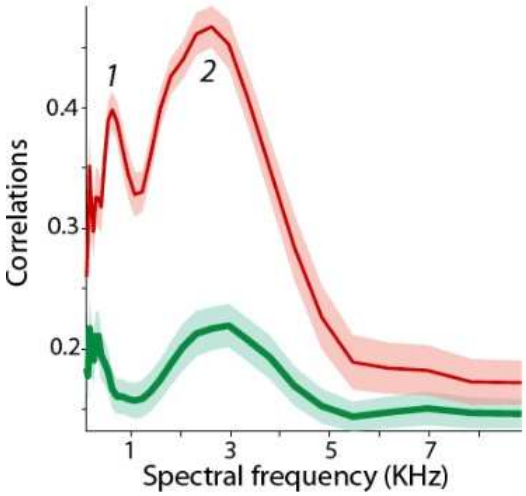


Fig. 7.08. Corrélation entre les signaux audio et visuel dans le domaine spectral [Chandrasekaran, 2009].

7.3.3.c. Structure temporelle de la parole audiovisuelle

Les spectres fréquentiels de l'enveloppe du signal audio et de la fonction temporelle de l'aire de la bouche sont calculés. Pour les deux spectres, l'allure est en 1/f et on observe un pic dans la bande 2–6 Hz (cf. Fig. 7.09). Or, cet intervalle est très similaire aux fréquences reliées aux oscillations Theta (3-8 Hz) du cerveau, qui sont impliquées dans le traitement de la parole.

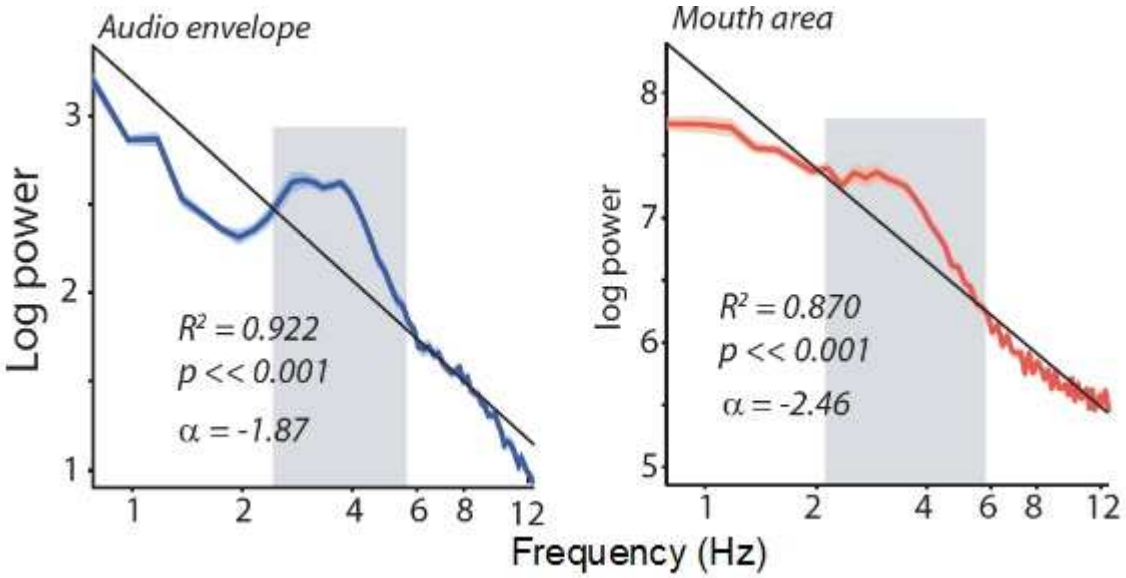


Fig. 7.09. Corrélation entre les spectres fréquentiels audio et visuel [Chandrasekaran, 2009].

7.3.4. Conclusion de l'étude

Dans [Chandrasekaran, 2009], les auteurs montrent la forte corrélation qui existe entre les signaux audio et vidéo. L'information visuelle a été caractérisée par l'algorithme de suivi des contours des lèvres proposé dans cette thèse. Cette étude met en évidence le fait que le signal visuel est une information complémentaire intéressante pour aider à la perception de la parole. La même étude a été réalisée avec les bases d'images Wisconsin x-ray microbeam database (langue anglaise,[Westbury, 1994]) et LIUM-AVS database (langue française, [Daubias, 2003]). Dans un soucis de comparaison avec une vérité de terrain, l'extraction des paramètres labiaux a été facilitée sur ces deux bases de données par la présence d'artifices (points annotés et maquillage bleu). Les résultats montrent les mêmes corrélations.

CONCLUSION ET PERSPECTIVES

Dans ce manuscrit, nous avons présenté une méthode pour segmenter automatiquement les contours extérieur et intérieur des lèvres. L'algorithme proposé est composé de deux modules, un module statique pour extraire les contours dans une image statique et un module de suivi pour suivre les contours dans des séquences vidéos. La méthode consiste à combiner les contours actifs et les modèles paramétriques afin de tirer profit des avantages de chacun et ainsi développer un algorithme précis et robuste. Ce choix a été motivé par les contraintes associées à nos deux applications cibles : le logiciel Makeuponline et le projet TELMA.

L'approche contour choisie a nécessité le développement de plusieurs gradients spécifiques combinant des informations de chrominance et de luminance pour accentuer le contour des lèvres. Nous avons notamment employé deux pseudo-teintes ([Poggio, 1998] [Liévin, 2004]) qui augmentent efficacement le contraste entre les pixels lèvre et les pixels peau.
Pour modéliser le contour labial, nous avons complété le modèle paramétrique du contour extérieur introduit dans [Eveno, 2003] avec un modèle interne qui peut prendre deux formes différentes selon que la bouche est ouverte ou fermée. Les modèles classiques étant trop simples (composés de paraboles) ou trop rigides (contraintes de symétrie), nous avons utilisé des courbes cubiques appropriées aux déformations possibles des lèvres.
Dans le cas d'une image statique, les deux premières phases du module statique sont la localisation du visage et la détection d'une boîte englobante de la bouche pour définir la région d'intérêt. Ensuite, les contours extérieur et intérieur des lèvres sont extraits automatiquement en combinant contours actifs et modèles paramétriques. Les jumping snakes permettent de trouver des points clefs externes et internes sur les contours afin de positionner le modèle paramétrique. La convergence des snakes et l'optimisation du modèle paramétrique des lèvres sont obtenues à partir d'une méthode de maximisation de flux moyens de gradients.
Dans le cas de séquences vidéo, le traitement statique est réalisé sur la première image pour initialiser l'algorithme de suivi. Les suivis de la boîte autour de la bouche et des points clefs externes et internes permettent d'initialiser efficacement et rapidement le modèle paramétrique des lèvres. Une mesure sur les déplacements des points clefs externes d'image en image rend l'algorithme robuste vis-à-vis de conditions difficiles de suivi (occultation partielle de la bouche par exemple) en permettant une réinitialisation du processus. Enfin, l'ajustement du modèle paramétrique s'effectue de nouveau par maximisation de flux moyen de gradients. La prise en compte d'informations temporelles permet d'accroître la vitesse de segmentation d'un facteur égal à 3,4 par rapport à la méthode statique.
Une évaluation quantitative de l'algorithme statique et de l'algorithme de suivi met en évidence la précision des résultats de segmentation obtenus. Une évaluation applicative permet, quant à elle, de montrer que la méthode proposée est appropriée pour les applications de mise en beauté virtuelle, en permettant notamment de ne maquiller que les lèvres lorsque la bouche est ouverte, et de lecture labiale.

A court terme, nous pouvons proposer plusieurs pistes qui permettraient d'améliorer la méthode de segmentation proposée.
Nous avons vu dans le chapitre 5 que l'algorithme statique éprouvait certaines difficultés à extraire le contour intérieur en présence de la langue. Il serait intéressant de voir s'il est possible de détecter la visibilité de la langue sur l'image avant la segmentation et modifier le traitement en fonction du résultat. La tâche s'annonce ardue dans la mesure où de manière générale, la langue et les lèvres ont des aspects similaires en termes de couleur et de texture.
Pour l'algorithme de suivi, des efforts restent encore à faire pour obtenir des résultats plus précis. Notamment, il serait nécessaire d'améliorer la détection de l'état de la bouche car il subsiste des cas difficiles lorsque la bouche est faiblement ouverte ou fermée mais avec une zone sombre importante au centre de la bouche.
La détection des commissures internes pourrait également être plus précise. Une solution pourrait être de coupler notre approche contour avec une approche région. Un critère région pourrait être défini afin de séparer au mieux la région des lèvres (définie par le contour extérieur moins la zone interne) et la région interne (définie par le contour intérieur).

Enfin, la zone du visage n'est détectée que dans la première image de la séquence avec l'algorithme C3F. Ensuite, on suppose que le visage bouge relativement peu pour le reste de la séquence. Pour que le locuteur soit plus libre de ses mouvements, il serait intéressant de suivre le cadre du visage en utilisant par exemple un filtre de Kalman.

A long terme, il serait envisageable de lever l'hypothèse de départ sur la pose du visage. En effet, tout au long de notre étude, nous avons traité des images où le visage était vu de face (ce qui est en accord avec l'application de Vesalis et le projet TELMA). Dans le cadre du projet TELMA, nous avons constaté que la méthode était suffisamment robuste vis-à-vis d'inclinaisons relativement faibles de la tête (Angle *roll*, qui correspond au mouvement de la tête qui se penche vers la gauche ou vers la droite, n'excédant pas ± 10°, cf. Fig. 6.23) mais nous n'avons pas testé les performances de l'algorithme pour des inclinaisons plus fortes ou lorsque la rotation de la tête s'effectue dans le plan de l'image (Angle *pan* qui correspond au mouvement de la tête qui se tourne de la gauche vers la droite). De la même façon, les séquences d'images que nous avons utilisées (base TELMA) pour tester l'algorithme de suivi proposé ont été acquises avec des conditions d'illumination constantes. Il serait intéressant de voir comment évolue le suivi des contours des lèvres lorsque ces conditions changent au cours d'une même séquence.

Concernant les applications cibles de notre travail, plusieurs perspectives sont envisageables à plus ou moins long terme.

La première étape consiste à implémenter l'algorithme statique en langage C pour l'intégrer au logiciel Makeuponline et maquiller les lèvres lorsque la bouche est ouverte. A plus long terme, l'implémentation de l'algorithme de suivi est prévu pour faire de la mise en beauté pour des visages en mouvement. D'autres algorithmes de suivi pour les autres traits permanents (contours du visage, des yeux et des sourcils) devront également être développés pour un maquillage complet. L'implémentation en langage C permettra également de voir si l'algorithme de suivi peut fonctionner en temps réel.

En outre, une amélioration importante concerne le cas des personnes ayant la peau noire. Le produit Makeuponline étant destiné à une commercialisation mondiale, et notamment américaine, il est impératif que la segmentation des lèvres soit possible quelque soit la couleur de la peau. Nous avons vu que l'algorithme proposé pouvait être mis en difficulté lorsque la peau et les lèvres ont une couleur similaire. Nous avons proposé des améliorations, mais cette étude doit être menée sur un plus grand nombre d'images. Il serait intéressant de disposer ou de créer une base d'images suffisamment importante.

Enfin, dans le cadre du projet TELMA, nous avons pu vérifier que l'algorithme fournissait des segmentations suffisamment précises pour obtenir des résultats en adéquation avec les études réalisées avec des artifices (maquillage bleu). Ceci a été fait pour caractériser les voyelles en fonction de la forme des lèvres et des paramètres labiaux internes. L'étape suivante serait de faire de la reconnaissance de phonèmes, c'est-à-dire reconnaître les voyelles dans un message à partir des paramètres labiaux obtenus à partir du signal vidéo.

REFERENCES

[Abboud, 2004] B. Abboud and F. Davoine, Facial Expression Recognition and Synthesis based on an Appearance Model, *Signal Processing: Image Communication*, Elsevier, vol. 19, no. 8, pp. 723-740, September 2004.

[Abboud, 2005] B. Abboud and G. Chollet, Appearance based Lip Tracking and Cloning on Speaking Faces, *in Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA 2005)*, pp. 301-305, September 2005.

[Aboutabit, 2007a] N. Aboutabit, Reconnaissance de la Langue française Parlée Complétée (LPC) : Décodage phonétique des gestes main-lèvres, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, décembre 2007.

[Aboutabit, 2007b] N. Aboutabit, D. Beautemps, J. Clarke, and L. Besacier, A HMM Recognition of Consonant-vowel Syllables from Lip Contours: the Cued Speech Case, *in Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[Ahlberg, 2001] J. Ahlberg, CANDIDE 3 - An Updated Parameterized Face, *Report No. LiTH-ISY-R-2326*, Dept. of Electrical Engineering, Linköping University, Sweden, 2001.

[Ballerini, 1999] L. Ballerini, Genetic Snakes for Medical Images Segmentation, *Lecture Notes in Computer Science*, vol. 1596, pp. 59-73, 1999.

[Barnard, 2002] M. Barnard, E. J. Holden and R. Owens, Lip Tracking using Pattern Matching Snakes, *The 5th Asian Conference on Computer Vision (ACCV'2002)*, pp. 273-278, January 2002.

[Baron, 1994] L. Barron, S. S. Beauchemin, and D. J. Fleet, Performance of Optical Flow Techniques, *In International Journal on Computer Vision*, vol. 12, pp. 43-77, 1994.

[Beaumesnil, 2006] B. Beaumesnil, M. Chaumont and F. Luthon, Liptracking and MPEG4 Animation with Feedback Control, *IEEE International Conference On Acoustics, speech, and Signal Processing, (ICASSP'2006)*, May 2006.

[Benoît, 1992] C. Benoit, T. Lallouache, T. Mohamadi and C. Abry, Talking Machines : Theories, Models and Designs, C*hapter A Set of French Visemes for Visual French Speech Synthesis*, Elsevier SC. Publishers, pp. 485-504, Amsterdam, 1992.

[Benoît, 1994] C. Benoît, T. Mohamadi and S. Kandel, Effects of Phonetic Context on Audio-Visual Intelligibility of French, *J. Speech and Hearing Research*, pp. 1195-1293, 1994.

[Benoît, 1996] C. Benoît, T. Guiard-Marigny, B. Le Goff and A. Adjoudani, Which Components of the Face Humans and Machines best Speechread? *Speechreading by Man and Machine: Models, Systems and Applications, Springer-Verlag*, D. G. Stork and M. E. Hennecke editors, pp. 315-328, New-York, 1996.

[Beskow, 1997] J. Beskow, M. Dahlquist, B. Granström M. Lundeberg, K. E. Spens and T. Öhman, The Teleface Project - Multimodal Speech Communication for the Hearing Impaired, *in Proceedings of*

Eurospeech '97, Rhodos, Greece, 1997.

[Blake, 1995] A. Blake, M. A. Isard and D. Reynard, Learning to Track the Visual Motion of Contours, *Artificial Intelligence*, pp. 101-134, 1995.

[Botino, 2002] A. Botino, Real Time Head and Facial Features Tracking from Uncalibrated Monocular Views. *In Proceedings 5th Asian Conference on Computer Vision (ACCV'02)*, Melbourne, Australia, 23-25 January 2002.

[Bouvier, 2007] C. Bouvier, P.-Y. Coulon and X. Maldague, Unsupervised Lips Segmentation Based on ROI Optimisation and Parametric Model, *International Conference on Image Processing*, (ICIP07), 2007.

[Brand, 2001] J. D. Brand, Visual Speech for Speaker Recognition and Robust Face Detection, *PhD thesis*, University of Wales, UK, 2001.

[Brunelli, 1995] R. Brunelli and D. Falavigna, Person Identification Using Multiple Cues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, October 1995.

[Brugger, 2006] F. Brugger, L. Zouari, H. Bredin, A. Amehraye, G. Chollet, D. Pastor and Y. Ni, Reconnaissance Audiovisuelle de la Parole par VMike, *XXVIes Journées d'Etude sur la Parole*, Manoir de la Vicomté, Dinard, France, http://jep2006.irisa.fr/, 12-16 juin 2006.

[Burger, 2007] T. Burger, Reconnaissance automatique des gestes de la Langue française Parlée Complétée, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, octobre 2007.

[Caplier, 2009] A. Caplier, S. Stillitano, C. Bouvier, P.-Y. Coulon, Chapter III : Lip Modeling and Segmentation, *Visual Speech Recognition : Lip Segmentation and Mapping*, A. Liew and S. Wang editors, IGI Global, 2009.

[Castañeda, 2005] B. Castañeda and J. C. Cockburn, Reduced Support Vector Machine Applied to Real-time Face Tracking, In Proceedings ICASSP, vol. 2, pp. 673-676, Philadelphia, 2005.

[Chan, 1998] M. T. Chan, Y. Zhang and T. S. Huang, Real-Time Lip Tracking and Bimodal Continuous Speech Recognition. *In Proceedings IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 65–70, Los Angeles, Calif, USA, December 1998.

[Chan, 2001] M. T. Chan, Hmm-based Audio-Visual Speech Recognition Integrating Geometric and Appearance-based Visual Features. *In Workshop on Multimedia Signal Processing*, 2001.

[Chandrasekaran, 2009] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier and A. A. Ghazanfar, Natural Statistics of Audiovisual Speech, *journal PLOS computational biology,* 2009 (proposé).

[Chen, 2004] S. C. Chen, C. L. Shao, C. K. Liang, S. W. Lin, T. H. Huang, M. C. Hsieh, C. H. Yang, C. H. Luo and C. M. Wuo, A Text Input System Developed by using Lips Image Recognition based LabVIEW for the seriously disabled, *International Conference of the IEEE Engineering in Medicine and Biology Society* (IEMBS'2004), vol. 2, pp. 4940-4943, 2004.

[Chen, 2006] Q. C. Chen, G. H. Deng, X. L. Wang and H. J. Huang, An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech, *International conference on Machine Learning and Cybernetics*, pp. 3859-3864, 2006.

[Chibelushi, 1997] C. Chibelushi, Automatic Audio-Visual Person Recognition, *PhD thesis*, University of Wales, Swansea, 1997.

[Chiou, 1997] G. Chiou and J. Hwang, Lipreading from Color Video. *In Trans. on Image Processing*, vol. 6(8), pp. 1192-1195, August 1997.

[Cohen, 1991] L. Cohen, On Active Contour Models and Balloons, *CVGIP: Image Understanding*, vol. 53(2), pp. 211-218, 1991.

[Cohn, 1998] J. Cohn, A. Zlochower, J.J. Lien and T. Kanade, Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression, *In International Conference on Automatic Face and* Gesture Recognition, pp. 396–401, Nara, Japan, 1998.

[Coianiz, 1996] T. Coianiz, L. Torresani and B. Caprile, 2D Deformable Models for Visual Speech Analysis, *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork & M. E. Hennecke editors, Springer-Verlag, pp. 391-398, New York, 1996.

[Cooke, 2006] M. Cooke, J. Barker, S. Cunningham, and X. Shao, An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition, *Journal of the Acoustical Society of America*, pp. 2421-2424, 2006.

[Cootes, 1992] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, Training Models of Shape from Sets of Examples, *3rd British Machine Vision Conference*, D. Hogg and R. Boyle editors, Springer-Verlag, pp. 9-18, September 1992.

[Cootes, 1995] T. F. Cootes, C. J. Taylor and D. H. Cooper, Active Shape Models - Their Training and Application, *Computer Vision and Image Understanding*, vol. 61, No. 1, pp. 38-59, 1995.

[Cootes, 1998] T. F. Cootes, G. J. Edwards and C. J. Taylor, Active Appearance Model, *In Proceedings European Conference on Computer Vision*, vol. 2, pp. 484-498, 1998.

[Cootes, 2004] T. F. Cootes, Statistical Models of Appearance for Computer Vision, *Technical report*, free to download on http://www.isbe.man.ac.uk/bim/refs.html, 2004.

[Daubias, 2002] P. Daubias and P. Deleglise, Statistical Lip-Appearance Models Trained Automatically using Audio Information, *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1202-1212, 2002.

[Daubias, 2003] P. Daubias ans P. Deleglise, The LIUM-AVS Database: A Corpus to Test Lip Segmentation and Speechreading Systems in Natural Conditions, *Proceedings 8th Eur. Conference on Speech Communication and Technology*, pp. 1569-1572, Geneva, Switzerland, September 2003.

[Deligne, 2002] S. Deligne, G. Potamianos and C. Neti, Audio-Visual Speech Enhancement with AVCDCN (Audio Visual Codebook Dependent Cepstral Normalization), *In Proceedings of International Conference Spoken Language Processing (ICSLP)*, pp. 1449-1452, 2002.

[Delmas, 1999] P. Delmas, P.-Y., Coulon and V. Fristot, Automatic Snakes for Robust Lip Boundaries Extraction, *International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3069-3072, 1999.

[Delmas, 2000] P. Delmas, Extraction des contours de lèvres d'un visage parlant par contours actifs, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, 2000.

[Delmas, 2002] P. Delmas, N. Eveno and M. Liévin, Towards Robust Lip Tracking, *International Conference on Pattern Recognition (ICPR02)*, vol. 2, pp. 528-531, 2002.

[Dornaika, 2004] F. Dornaika and F. Davoine, Head and Facial Animation Tracking using Appearance-adaptive Models and Particle Filters, *IEEE CVPR Workshop on Real-time vision for human computer*

interaction, Washington, U.S.A., July, 2004.

[Duffner, 2005] S. Duffner and C. Garcia, A Hierarchical Approach for Precise Facial Feature Detection, *Compression et Representation des Signaux Audiovisuels (CORESA)*, Rennes, France,2005.

[Easton, 1982] R.D. Easton and M. Basala, Perceptual Dominance during Lipreading, *Perception and Psychophysics*, vol. 32, pp 562-570, 1982.

[Ekman, 1978] P. Ekman and W. V. Friesen, Facial Action Coding System (FACS): Manual, Palo Alto : Consulting Psychologists Press, 1978.

[Ekman, 1982] P. Ekman, Emotion in the Human Face, *Cambridge University Press*, 1982.

[Essa, 1994] I.A. Essa and A. Pentland, A Vision System for Observing and Extracting Facial Action Parameters, *In Proceedings of Computer Vision and Pattern Recognition (CVPR 94)*, pp. 76-83, 1994.

[Essa, 1995] I.A. Essa and A.P. Pentland, Facial Expression Recognition using a Dynamic Model and Motion Energy, *Fifth International Conference on Computer Vision (ICCV'95)*, pp. 360, 1995.

[Eveno, 2003] N. Eveno, Segmentation des lèvres par un modèle déformable analytique, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, novembre 2003.

[Eveno, 2004] N. Eveno, A. Caplier and P. Y. Coulon, Automatic and Accurate Lip Tracking, *IEEE Transactions on Circuits and Systems for Video technology*, vol. 14, no. 5, pp.706-715, May 2004.

[Finn, 1988] K.E. Finn and A.A. Montgomery, Automatic Optically-based recognition of speech, *Pattern Recognition Letters*, vol. 8, no. 3, pp. 159-164, 1988.

[Fleet, 1990] D. J. Fleet and A. D. Jepson, Computation of Component Image Velocity from Local Phase Information, *In International Journal on Computer Vision*, vol. 5, pp. 77-104, 1990.

[Ford, 1998] A. Ford and A. Roberts, *Color Space Conversion* (Technical Report), http://inforamep.net/poyton/PDFs/coloureq.pdf, 1998.

[Gacon, 2005] P. Gacon, P.-Y. Coulon, and G. Bailly, Non-Linear Active Model for Mouth Inner and Outer Contours Detection, *Proceedings of European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turkey, 2005.

[Gacon, 2006] P. Gacon, Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, juillet 2006.

[Garcia, 2004] C. Garcia adn M. Delakis, Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1408-1423, 2004.

[Girin, 1997] L. Girin, G. Feng and J. L. Schwartz, Débruitage de Parole par Fusion des Informations Auditives et Visuelles : une Etude des Transitions Vocalique, *Congrès GRETSI 97 : seizième Colloque sur le Traitement du Signal et des Images*, pp. 1407-1410, Grenoble, FRANCE, 1997.

[Girin, 2004] L. Girin, Joint Matrix Quantization of Face Parameters and LPC Coefficients for Low Bit Rate Audiovisual Speech Coding, *IEEE Transactions on Speech and Audio Processing*, vol. 12(3), pp. 265-276, May 2004.

[Goecke, 2002] R. Goecke, G. Potamianos and C. Neti, Noisy Audio Feature Enhancement using Audio-Visual Speech Data, *In Proceedings International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2025-2028, Orlando, USA, May 2002.

[Goldshen, 1993] A.J. Goldschen, Continuous Automatic Speech Recognition by Lipreading, *PhD dissertation*, George Washington University, September 1993.

[Gong, 1995] Y. Gong and M. Sakauchi, Detection of Regions Matching Specified Chromatic Features, *Computer Vision and Image Understanding*, vol. 61, pp. 263-269, 1995.

[Hammal, 2007] Z. Hammal, L. Couvreur, A. Caplier and M. Rombaut, Facial Expression Classification: An Approach based on the Fusion of Facial Deformation using the Transferable Belief Model, *International Journal of Approximate Reasoning*, vol. 46, pp. 542-567, December 2007.

[Heckmann, 2002] M. Heckmann, K. Kroschel, C. Savariaux and F. Berthommier, DCT-based Video Features for Audio-Visual Speech Recognition, *In Proceedings of the 7th ICSLP*, vol. 3, pp. 1925-1928, Denver, Colorado (USA), 2002.

[Hennecke, 1994] M. Hennecke, K. Prasad and Stork, D, Using Deformable Template to Infer Visual Speech Dynamics, *In Proceedings 28th Annual Asilomar Conference on Signal, Systems and Computers*, pp.578-582, 1994.

[Hong, 2006] T. Hong, Y. B. Lee, Y. G. Kim and H. Kim, Facial Expression Recognition using Active Appearance Model, *Third International Symposium on Neural Networks ISNN*, vol. 3973, pp. 69-76, Chengdu, China, May-June 2006.

[Horbelt, 1995] S. Horbelt and J. L. Dugelay, Active Contours for Lipreading – Combining Snakes with Templates, *GRETSI symposium on Signal and Image Processing*, France, 1995.

[Jang, 2006] K. S. Jang, S. Han, I. Lee and Y. W. Woo, Lip Localization Based on Active Shape Model and Gaussian Mixture Model, *In Advances in Image and Video Technology*, Springer Berlin / Heidelberg, vol. 4319, pp.1049-1058, 2006.

[Jian, 2001] Z. Jian, M. N. Kaynak, A. D. Cheok and K. C. Chung, Real-Time Lip Tracking for Virtual Lip Implementation in Virtual Environments and Computer Games, *International Conference on Fuzzy Systems*, vol. 3, pp. 1359- 1362, 2001.

[Jian, 2006] Y.-D. Jian, W.-Y. Chang and C.-S. Chen, Attractor-Guided Particle Filtering for Lip Contour Tracking, *In Asian Conference on Computer Vision*, vol. 1, pp. 653-663, 2006.

[Jing, 2000] Z. Jing, R. Mariani and J. Wu, Glasses Detection for Face Recognition Using Bayes Rules, *International Conference on Multimodal Interfaces (ICMI 2000)*, vol. 1948, pp. 127-134, Beijing, China, October 2000.

[Jourlin, 1997] P. Jourlin, J. Luettin, D. Genoud and H. Wassber, Acoustic Labial Speaker Verification, *In Proceedings AVBPA, Lecture Notes in Computer Science* 1206, pp. 319-334, 1997.

[Kalman, 1960] R. E. Kalman, A New Approach to Linear Filtering and Prediction Problems, *Transaction of the ASME - Journal of Basic Engineering*, vol. 82, pp. 35-45, 1960.

[Kass 1987] M. Kass, A. Witkin and D. Terzopoulos, Snakes: Active Contour Models, *International Journal of Computer Vision*, vol. 1(4), pp. 321-331, 1987.

[Kaucic, 1998] R. Kaucic and A. Blake, Accurate, Real-Time, Unadorned Lip Tracking, *International*

Conference on Computer Vision, pp. 370-375, 1998.

[Kim, 2005] S. M. Kim, K. C. Seo and S. W. Lee, Image-Based Generation of Facial Skin Texture with Make-Up, *Lecture Notes in Computer Science* 3767, pp. 350-360, 2005.

[Kuo, 2005] P. Kuo, P. Hillman and J. M. Hannah, Improved Lip Fitting and Tracking for Model-based Multimedia and Coding, *Visual ,Information Engineering*, pp. 251-258, 2005.

[Lallouache, 1991] T. Lallouache, Un poste Visage-Parole. Acquisition et traitement automatique des contours des lèvres, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, 1991.

[Lanitis, 1997] A. Lanitis, C.J. Taylor and T. F. Cootes, Automatic Interpretation and Coding of Face Images using Flexible Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pp. 743-756, July 1997.

[Lanitis, 2002] A. Lanitis, C. J. Taylor and T. F. Cootes, Toward Automatic Simulation of Aging Effects on Face Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442-445, April 2002.

[Le Gallou, 2006] S. Le Gallou, G. Breton, C. Garcia, R. Séguier., Cartes de distances : prétraitement robuste à l'illumination pour les modèles actifs d'apparence (AAM), *MajecStic* Lorient, France, novembre 2006.

[Lehn-Schiøler, 2004] T. Lehn-Schiøler, Multimedia Mapping using Continuous State Space Models, *IEEE 6th Workshop on Multimedia Signal Processing Proceedings*, pp. 51-54, 2004.

[Leung, 2004] S.-H. Leung, S.-L. Wang and W.-H. Lau, Lip Image Segmentation Using Fuzzy Clustering Incorporating an Elliptic Shape Function, *IEEE Transactions on Image Processing*, vol. 13, No. 1, pp. 51-62, 2004.

[Liévin, 2004] M. Liévin and F. Luthon, Nonlinear Color Space and Spatiotemporal MRF for Hierarchical Segmentation of Face Features in Video, *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 63-71, January 2004.

[Liew, 2000] A. Liew, S. H. Leung and W. H. Lau, Lip Contour Extraction using a Deformable Model, *International Conference on Image Processing*, vol. 2, pp. 255-258, 2000.

[Liew, 2003] A. Liew, S. H. Leung and W. H. Lau, Segmentation of Color Lip Images by Spatial Fuzzy Clustering, *IEEE Transactions on Fuzzy Systems*, vol. 11, No. 1, pp. 542-549, 2003.

[Li, 2006] Z. Li and H. Ai, Texture-Constrained Shape Prediction for Mouth Contour Extraction and its State Estimation, *In Proceedings ICPR06*, vol. 2, pp. 88-91, Hong Kong, 2006.

[Liu, 2006] W. Liu and Z. F. Wang, Facial Expression Recognition based on Fusion of Multiple Gabor Features, *18th International Conference on Pattern Recognition (ICPR 2006)*, vol. 3, pp. 536-539, 2006.

[Lucas, 1981] B. D. Lucas and T. Kanade, An Iterative Image Registration Technique with an Application to Stereo Vision. *In Proceedings IJCAI'81*, pp. 674-679, Vancouver, Canada, 1981.

[Lucas, 1984] B. D. Lucas, Generalized Image Matching by the Method of Differences, *Carnegie Mellon University*, Technical Report CMU-CS-85-160, Ph.D. dissertation, July 1984.

[Lucey, 1999] S. Lucey, S. Sridharan and V. Chandran, Chromatic Lip Tracking using a Connectivity based Fuzzy Thresholding Technique, *In ISSPA99*, 1999.

[Luettin, 1996] J. Luettin, N. A. Thacker and S. W. Beet, Speaker Identification by Lipreading, *in 4th International Conference on Spoken Language (ICSLP 96)*, vol. 1, pp. 62-65, 1996.

[Luettin, 1997] J. Luettin, Visual Speech and Speaker Recognition, *Ph.D. thesis*, University of Sheffield, Sheffield, UK, 1997.

[Lyons, 1998] M. Lyons and S. Akamatsu, Coding Facial Expressions with Gabor Wavelets Proceedings, *Third IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society, pp. 200-205, Nara, Japan, April 1998.

[Lyons, 2003] M. J. Lyons, M. Haehnel and N. Tetsutani, Designing, Playing, and Performing with a Vision-Based Mouth Interface, *In Proceedings Conference on New Interfaces for Musical Expression (NIME-03)*, pp. 116-121, 2003.

[Malciu, 2000] M. Malciu and F. Prêteux, Tracking Facial Features in Video Sequences Using a Deformable Model-based approach, *Proceedings SPIE Mathematical Modeling, Estimation, and Imaging*, vol. 4121, pp. 51-62, 2000.

[Martinez, 1998] A. M. Martinez and R. Benavente, The AR Face Database, *CVC Technical Report* #24, 1998.

[Mase, 1991] K. Mase and A. Pentland, Automatic Lipreading by Optical Flow Analysis, *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67-75, 1991.

[Matthews, 1998] I. Matthews, T. F. Cootes, S. Cox, R. Harvey and J. A. Bangham, Lipreading using Shape, Shading and Scale, *In Proceedings Auditory-Visual Speech Processing (AVSP)*, pp.73-78, Australia, 1998.

[Matthews, 2003] I. Matthews, S. Baker S, Active Appearance Models Revisited, *Technical Report CMU-RITR-03-02*, Carnegie Mellon University Robotics Institute, free to download on http://citeseer.ist.psu.edu/matthews04active.html, April 2003.

[McGurk, 1976] H. McGurk and J. McDonald, Hearing Lips and Seeing Voices, *Nature*, pp. 746-748, December 1976.

[Mirhosseini, 1997] A. R. Mirhosseini, C. Chen, K. M. Lam and H. Yan, A Hierarchical and Adaptive Deformable Model for Mouth Boundary Detection, *International Conference on Image Processing*, vol. 2, pp. 756-759, 1997.

[Morishima, 1989] S. Morishima, K. Aizawa and H. Harashima, An Intelligent Facial Image Coding Driven by Speech and Phoneme, *In Proceedings IEEE ICASSP*, pp. 1795, Glasgow, UK, 1989.

[Morishima, 1995] S. Morishima, Emotion model, *International Workshop on Automatic Face and Gesture Recognition*, pp. 284-289, Zurich, Switzerland, 1995.

[MPEG, 2001] MPEGWorking Group on VISUAL, International Standard on Coding of Audio-Visual Objects, Part 2 (Visual), 2001. ISO/IEC 14496-2 :2001.

[Muhammad Hanif, 2007] S. Muhammad Hanif, L. Prevost, R. Belaroussi and M. Milgram, Real-time facial feature localization by combining space displacement neural networks, *Pattern Recognition Letters, Special issue on Pattern Recognition in Multidisciplinary Perception and Intelligence*, vol. 29(8), pp. 1094-1104, 2007.

[Neely, 1956] K. K. Neely, Effect of Visual Factors on the Intelligibility of Speech, *J. Acoustical Society of America.* Vol. 28, pp. 1275-1277, 1956.

[Nefian, 2002] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao and K. Murphy, A Coupled HMM for Audio-Visual Speech Recognition, *In Proceedings 2002 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 2013-2016, Orlando, 2002.

[Nkambou, 2004] R. Nkambou and V. Heritier, Reconnaissance Emotionnelle par l'Analyse des Expressions Faciales dans un Tuteur Intelligent Affectif, *Technologies de l'Information et de la Connaissance dans l'Enseignement Supérieur et l'Industrie*, pp. 149-155, France, 2004.

[Nishida, 1986] S. Nishida. Speech Recognition Enhancement by Lip Information, *ACM SIG-CHI Bulletin*, vol. 17 (4), pp. 198-204, 1986.

[Olivier, 2000] N. Oliver, A. Penland and F. Bérard, LAFTER: a Real-Time Face and Lips Tracker with Facial Expression Recognition, *Pattern Recognition*, vol. 33, pp. 1369-1382, 2000.

[Otsu, 1979] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, *in IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62-66, 1979.

[Pantic, 2001] M. Pantic, M. Tomc and L. J. M. Rothkrantz, A hybrid Approach to Mouth Features Detection, *Proceedings of IEEE International Conference Systems, Man and Cybernetics (SMC'01),* pp. 1188-1193, Tucson, USA, October 2001.

[Pardas, 2001] M. Pardàs and E. Sayrol, Motion Estimation Based Tracking of Active Contours, *Pattern Recognition Letters*, vol. 22, pp. 1447-1456, 2001.

[Park, 2005] J. S. Park, Y. H. Oh, S. C. Ahn and S. W. Lee, Glasses Removal from Facial Image using Recursive Error Compensation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, Issue 5, pp. 805-811, May, 2005.

[Patterson, 2002] E. K. Patterson, S. Gurbuz, Z. Tufekci and J. H. Gowdy, Moving-Talker, Speaker-Independent Feature Study and Baseline Results using the CUAVE Multimodal Speech Corpus, *EURASIP Journal on Applied Signal Processing*, Issue 11, pp. 1189-1201, 2002.

[Petajan, 1984] E. Petajan, Automatic Lipreading to Enhance Speech Recognition, *PhD thesis*, University of Illinois at Urbana-Champaign, 1984.

[Phillips, 1998] P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, The FERET Database and Evaluation Procedure for Face Recognition Algorithms, *Image and Vision Computing J*, vol. 16, no. 5, pp. 295-306, 1998.

[Poggio, 1998] T. Poggio and A. Hulbert, Synthesizing a Color Algorithm from Examples, *Science*, vol. 239, pp. 482-485, 1998.

[Poh, 2001] N. Poh and J. Korczak, Hybrid Biometric Person Authentication using Face and Voice Features, *Proceedings of International Conference on Audio and Video-Based Biometric Person Authentication*, pp.348-353, Halmstad, Sweden, June 2001.

[Potamianos, 1997] G. Potamianos, E. Cosatto, H.P. Graf and D.B.Roe, Speaker Independent Audio-Visual Database for Bimodal ASR, *In Proceedings of the European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, September 1997.

[Potamianos, 1998] G. Potamianos, H. P. Graf, and E. Cosatto, An Image Transform Approach for HMM

based Automatic Lipreading, *in* Proceedings of the International Conference on Image Processing (ICIP), Chicago, 1998.

[Potamianos, 2004] G. Potamianos, C. Neti, J. Luettin and I. Matthews, Audio-Visual Automatic Speech Recognition: an Overview, *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier editors, MIT Press., 2004.

[Radeva, 1995] P. Radeva E. Marti, Facial Features Segmentation by Model-Based Snakes, *International Conference on Computer Vision (ICCV95)*, 1995.

[Rao, 1995] R. Rao and R. Mersereau, On Merging Hidden Markov Models with Deformable Templates, *International Conference on Image Processing (ICIP'1995)*, vol. 3, pp. 556-559, 1995.

[Ratliff, 2008] M. Ratliff and E. Patterson, Emotion Recognition using Facial Expressions with Active Appearance Models, *Proceedings of the IASTED International Conference on Human-Computer Interface*, Innsbruck, March 2008.

[Rehman, 2007] S. U. Rehman, L. Liu and H. Li., Lip Localization and Performance Evaluation, *Proceedings IEEE International Conference on Machine Vision (ICMV'2007)*, pp. 29-34, Islamabad, Pakistan, 2007.

[Rivet, 2006a] B. Rivet, La bimodalité de la parole au secours de la séparation de sources, *Thèse de doctorat*, Institut National Polytechnique de Grenoble, septembre 2006.

[Rivet, 2006b] B. Rivet, C. Servière, L. Girin, D. T. Pham, and C. Jutten, Un Détecteur d'Activité Vocale Visuel pour Résoudre le Problème des Permutations en Séparation de Source de Parole dans un Mélange Convolutif, *In Proceedings Journées d'Etude sur la Parole (JEP)*, pp. 85-88, Dinard, France, juin 2006.

[Rivet, 2007] B. Rivet, A. Aubrey, L. Girin, Y. Hicks, C. Jutten, and J. Chambers, Development and Comparison of two Approaches for Visual Speech Analysis with Application to Voice Activity Detection, *In Proceedings International Conference Auditory-Visual Speech Processing (AVSP)*, pp. 228-232, Hilvarenbeek, The Netherlands, September 2007.

[Robert-Ribès, 1995] J. Robert-Ribès, Modèles d'intégration audiovisuelle de signaux linguistiques: de la perception humaine à la reconnaissance automatique des voyelles, Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 1995.

[Ross, 2004] A. Ross and A. K. Jain, Multimodal Biometrics: An Overview, *Proceedings of 12th European Signal Processing Conference (EUSIPCO)*, pp. 1221-1224, Vienna, Austria, September 2004.

[Rowland, 1995] D. A. Rowland and D. I. Perrett, Manipulating Facial Appearance through Shape and Color, *IEEE Computer Graphics and Applications*, vol. 15, no. 5, pp. 70-76, September 1995.

[Salazar, 2007] A. Salazar, J. E. Hernández and F. Prieto, Automatic Quantitative Mouth Shape Analysis, *International Conference on Computer Analysis of Images and Patterns (CAIP'2007)*, pp. 416-423, 2007.

[Sasaki, 2004] Y. Sasaki, T. Kawamura and K. Sugahara, Lip Shape Extraction for Word Recognition by using Hardware Active Contour Model, *International Symposium on Intelligent Multimedia, Video and Speech Processing*, pp. 370-373, 2004.

[Seguier, 2003] R. Seguier and N. Cladel, Genetic Snakes: Application on Lipreading, *International Conference on Artificial Neural Networks and Genetic Algorithms, (ICANNGA)*, 2003.

[Seo, 2003] K. H. Seo and J. J. Lee, Object Tracking using Adaptive Color Snake Model, *International*

Conference on Advanced Intelligent Mechatronics (AIM'2003), vol. 2, pp. 1406-1410, 2003.

[Seyedarabi, 2006] H. Seyedarabi, W. Lee and A. Aghagolzadeh, Automatic Lip Tracking and Action Units Classification using Two-step Active Contours and Probabilistic Neural Networks, *Canadian Conference on Electrical and Computer Engineering, (CCECE'2006)*, pp. 2021-2024, 2006.

[Shinchi, 1998] T. Shinchi, Y. Maeda, K. Sugahara and R. Konishi, Vowel Recognition According to Lip Shapes by using Neural Network, *In The IEEE International Joint Conference on Neural Networks Proceedings and IEEE World Congress on Computational Intelligence*, vol. 3, pp.1772-1777, 1998.

[Sodoyer, 2004] D. Sodoyer, L. Girin, C. Jutten and J. L. Schwartz, Developing an Audio-Visual Speech Source Separation Algorithm, *Speech Communication*, vol. 44 (1-4), pp. 113-125, October 2004.

[Stillittano, 2008] S. Stillittano and A. Caplier, Inner Lip Contour Segmentation by Combining Active Contours and Parametric Models, *International Conference on Computer Vision Theory and Applications (VISAPP 2008)*, Madeira, Portugal, January 2008.

[Sugahara, 2000] K. Sugahara, M. Kishino and R. Konishi, Personal Computer based Real Time Lip Reading System, *In Signal Processing Proceedings, WCCC-ICSP2000*, vol. 2, pp.1341-1346, 2000.

[Sumby, 1954] W. H. Sumby and I. Pollack, Visual Contribution to Speech Intelligibility in Noise, *J. Acoustical Society of America.* Vol. 26, pp. 212-215, 1954.

[Summerfield, 1987] Q. Summerfield, Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception, *In Hearing by Eye : The Psychology of Lipreading*, B. Dodd and R. Campbell editors, pp. 3-51, 1987.

[Suzuki, 1970] K. Suzuki and Y. Tsuchihashi, Personal Identification by Means of Lip Print, *Journal of Forensic Medicine*, vol. 17(2), pp. 52-57, 1970.

[Terzopoulos, 1993] D. Terzopoulos and K. Waters, Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569-579, June 1993.

[Thiran, 2007] J. P. Thiran, A. Valles, T. Drugman and M. Gurban, Définition et Sélection d'Attributs Visuels pour la Reconnaissance Audio-Visuelle de la Parole, *Traitement et Analyse de l'Information : Méthodes et Applications (TAIMA07)*, Hammamet, Tunisie, 2007.

[Tian, 2000a] Y. Tian, T. Kanade and J. Cohn, Robust Lip Tracking by Combining Shape, Color and Motion, *4th Asian Conference on Computer Vision, (ACCV'2000)*, January 2000.

[Tian, 2000b] Y. Tian, T. Kanade, and J. Cohn, Eye-State Action Unit Detection by Gabor Wavelets, *In Proceedings of* International Conference on Multi-modal Interfaces, pp. 143–150, September 2000.

[Tian, 2001] Y. Tian, T. Kanade and J. F. Cohn, Recognizing Action Units for Facial Expression Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 97-115, February 2001.

[Tomasi, 1991] C. Tomasi and T. Kanade, Detection and Tracking of Point Features, *Technical Report* CMU-CS-91-132, Carnegie Mellon University, 1991.

[Tsapatsoulis, 2002] N. Tsapatsoulis, Y. Avrithis and S. Kollias, Efficient Face Detection for Multimedia Applications, *International Conference on Image Processing* (ICIP00), TA07.11, Vancouver, Canada, September 2000.

[Turk, 1991] M. Turk and A. Pentland, Eigenfaces for Recognition, *Journal Cognitive Neuroscience*, vol. 3, pp.71-86, 1991.

[Vogt, 1996] M. Vogt, Fast Matching of a Dynamic Lip Model to Color Video Sequences Under Regular Illumination Conditions, *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke editors, vol. 150, pp. 399-407, 1996.

[Wakasugi, 2004] T. Wakasugi, M. Nishiura and K Fukui, Robust Lip Contour Extraction using Separability of Multi-Dimensional Distributions, *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'2004)*, pp. 415-420, May 2004.

[Wang, 2004a] J. Wang and C. X. Ling, Artificial Aging of Faces by Support Vector Machines, *in Advances in Artificial Intelligence*, vol. 3060 of *Lecture Notes in Computer Science*, pp. 499-503, Springer, Berlin, Germany, 2004.

[Wang, 2004b] S. L. Wang, W. H. Lau, S. H. Leung and H. Yan, A Real-time Automatic Lipreading System, *In ISCAS, IEEE International Symposium on Circuits and Systems*, Vol. 2, pp. 101-104, 2004.

[Wang, 2005] W. Wang, D. Cosker, Y. Hicks, S. Sanei and J. A. Chambers, Video Assisted Speech Source Separation, *In Proceedings IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[Wang, 2007] S. L. Wang, W. H. Lau, A. C. Liew and S. H. Leung, Robust Lip Region Segmentation for Lip Images with Complex Background, *PR(40)*, No. 12, pp. 3481-3491, December 2007.

[Wark, 1998] T. Wark, S. Sridharan and V Chandran,An Approach to Statistical Lip Modeling for Speaker Identification via Chromatic Feature Extraction, *In Proceedings 14th ICPR*, vol. 1, pp.123-125, Brisbane, Australia, 1998.

[Werda, 2007] S. Werda, W. Mahdi and A. B. Hamadou, Automatic Hybrid Approach for Lip POI Localization: Application for Lip-Reading System, *1st International Conference on Information and Communication Technology and Accessibility (ICTA'07)*, Hammamet, Tunisia, April 2007.

[Westbury, 1994] J. R. Westbury, G. Turner and J. Dembovski, X-Ray Microbeam Speech Production Database Users' Handbook, Waisman Center, University of Wisconsin, 1994.

[Wojdel, 2001] J. C. Wojdel and L. J. M. Rothkrantz, Using Aerial and Geometric Features in Automatic Lip-Reading, *In Proceedings 7th Eurospeech*, vol. 4, pp. 2463-2466, Aalborg, Denmark, 2001.

[Wu, 2002] Z. Wu, A. Z. Petar and A. K. Katsaggelos, Lip Tracking for MPEG-4 Facial Animation, *International Conference on Multimodal Interfaces (ICMI'02)*, Pittsburgh, PA, October 2002.

[Wyszecki, 1982] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, John Wiley & Sons, Inc., New York, New York, 2nd edition, 1982.

[Xu, 1998] C. Xu and J. L. Prince, Snakes, Shapes, and Gradient Vector Flow, *IEEE Transactions on Image Processing*, vol. 7, pp. 359-369, 1998.

[Yacoob, 1994] Y. Yacoob and L. Davis, Recognizing Human Facial Expression, *Technical Report CS-TR-3265*, University of Maryland, May 1994.

[Yamada, 1993] H. Yamada, Dimensions of Visual Information for Categorizing Facial Expressions, *Japanese Psychol. Res.*, vol. 35 (4), pp. 172-181, 1993.

[Yang, 1996] J. Yang and A. Waibel, A Real-Time Face Tracker, *In Proceedings of 3rd IEEE Workshop on Applications of Computer Vision*, pp. 142-147, Sarasota, USA, 1996.

[Yang, 2002] M.-H. Yang, D. J. Kriegman and N. Ahuja, Detecting Faces in Image: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, January 2002.

[Yang, 2005] D.-D. Yang and L.-W. Jin, Facial Expression Recognition with Pyramid Gabor Features and Complete Kernel Fisher Linear Discriminant Analysis, *International Journal of Information Technology*, vol. 11, no. 9, pp. 91-100, 2005.

[Yin, 2002] L. Yin and A. Basu, Color-Based Mouth Shape Tracking for Synthesizing Realistic Facial Expressions, *International Conference on Image Processing (ICIP 2002)*, pp. 161-164, September 2002.

[Yokogawa, 2007] Y. Yokogawa, N. Funabiki, T. Higashino, M. Oda and Y. Mori, A Proposal of Improved Lip Contour Extraction Method using Deformable Template Matching and its Application to Dental Treatment, *Systems and Computers in Japan*, vol.38, pp. 80-89, May 2007.

[Yuhas, 1989] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski, Integration of Acoustic and Visual Speech Signals using Neural Networks, *IEEE Communication Magazine*, pp. 65-71, November 1989.

[Yuille, 1992] A. Yuille, P. Hallinan and D. Cohen, Features Extraction from Faces using Deformable Templates, *International Journal of Computer Vision*, vol. 8(2), pp. 99-111, 1992.

[Zhang, 1997] L. Zhang, Estimation of the Mouth Features using Deformable Template, *International Conference on Image Procesing* (ICIP'1997), vol. 3, pp. 328-331, October 1997.

[Zhang, 2000] X. Zhang and R. M. Mersereau, Lip Feature Extraction Towards an Automatic Speechreading System, *In Proceedings International Conference on Image Processing*, Vancouver, British Columbia, Canada, 2000.

[Zhang, 2002] X. Zhang, C. Broun, R. Mersereau and M. Clements, Automatic Speechreading with Applications to Human-Computer Interfaces, *Eurasip Journal on Applied Signal Processing*, vol. 11, pp. 1228-1247, 2002.

[Zhao, 2003] W. Zhao, R. Chellappa, A. Rosenfeld and P.J. Phillips, Face Recognition: A Literature Survey, *ACM Computing Surveys*, pp. 399-458, 2003.

[Zhiming, 2002] W. Zhiming, C. Lianhong and A. Haizhou, A Dynamic Viseme Model for Personalizing a Talking Head, *International Conference on Signal Processing*, vol. 2, pp. 1015-1018, 2002.

[Zou, 2007] F. Zou, D. Yang, S. Li, J. Xu, and C. Zhou, Level Set Method based Tongue Segmentation in Traditional Chinese Medicine, *From Proceeding Graphics and Visualization in Engineering*, 2007.

PUBLICATIONS

Chapitre de livre

- A. Caplier, S. Stillitano, C. Bouvier, P.-Y. Coulon, Chapter III : Lip Modeling and Segmentation, *Visual Speech Recognition : Lip Segmentation and Mapping*, A. Liew and S. Wang editors, IGI Global ISBN: 9781605661865, 2009.

Papier journal

- A.Caplier, S. Stillittano, O. Aran, L. Akarun, G. Bailly, D. Beautemps, N. Aboutabit and T. Burger, Image and Video for hearing impaired people, Article de tutorial pour le numéro spécial "Image and video for disabilities" EURASIP Journal on Image and Video Processing, vol. 2007, Article ID45641, 14 pages, 2007.

- C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier and A. A. Ghazanfar, Natural Statistics of Audiovisual Speech, *journal PLOS computational biology,* 2009

Conférence internationale

- S. Stillittano and A. Caplier, Inner Lip Segmentation by Combining Active Contours and Parametric Models, *VISAPP 2008 - International Conference on Computer Vision Theory and Applications*, pp. 297-304, Madeira, Portugal, January 2008.

- S. Stillittano, V. Girondel and A. Caplier, Inner and Outer Lip Contour Tracking using Cubic Curve Parametric Models, *IEEE International Conference on Image Processing (ICIP09)*, 2009.

Colloque national

- S. Stillittano, A. Caplier, Segmentation du contour intérieur des lèvres en combinant contours actifs et modèles paramétriques, *CORESA (COmpression et REprésentation des Signaux Audiovisuels) 2007*, Montpellier, France, Novembre 2007.

RESUME

Ces dernières années, l'analyse des visages connaît un intérêt grandissant dans le domaine de la vision par ordinateur. Le visage est un vecteur d'information puissant de la communication entre être humains et il fournit des indications pertinentes sur l'identité d'une personne, sur son état émotionnel ou sur ce qu'elle dit. Le laboratoire GIPSA a mené de multiples études concernant le problème de la segmentation automatique des traits du visage pour des applications de type multimédia (réalité mixte, terminal téléphonique, interaction homme machine, interprétation de gestes de communication non verbal, simulateur de conduite interactif…). Des travaux ont porté sur la localisation de la tête dans une image, sur l'extraction des contours des yeux, des sourcils et de l'arc mandibulaire et, plus récemment, sur la segmentation des contours de la bouche.

Cette thèse présente un algorithme automatique de segmentation des contours intérieur et extérieur des lèvres utilisé pour des images statiques et des séquences vidéo. Ce système est composé de deux modules : un module statique et un module de suivi.

Dans le cas d'une image statique, après avoir localisé le visage et avoir calculé une boîte englobante de la bouche, l'algorithme statique permet d'extraire automatiquement le contour complet des lèvres en combinant contours actifs et modèles paramétriques. Les jumping snakes permettent de trouver des points clefs externes et internes sur les contours afin de positionner un modèle paramétrique composé de courbes cubiques appropriées aux déformations possibles des lèvres. Le modèle interne peut prendre deux formes différentes selon que la bouche soit ouverte ou fermée. Finalement, une méthode de maximisation de flux moyen de gradients optimise le modèle paramétrique.

Dans le cas de séquences vidéo, le même traitement statique est réalisé sur la 1ère image pour initialiser l'algorithme de suivi. La segmentation des contours dans les images suivantes se fait à l'aide de méthodes de tracking permettant le suivi des points clefs du modèle paramétrique des lèvres. L'ajustement du modèle paramétrique s'effectue ensuite de nouveau par maximisation de flux moyen de gradients.

Les contributions de cette thèse sont les suivantes: 1) Proposition d'un modèle paramétrique complet des lèvres suffisamment flexible pour reproduire un ensemble varié de formes possibles de la bouche 2) Création de plusieurs gradients combinant des informations de luminance et de chrominance adaptés à chaque partie du contour labial. 3) Évaluation quantitative et qualitative de l'algorithme de segmentation dans le cadre d'applications de maquillage virtuel et de lecture labiale.

Mots clefs : Contours extérieur et intérieur des lèvres, segmentation, suivi, contours actifs (jumping snakes), modèles paramétriques, courbes cubiques, détection des dents, détection de l'état de la bouche (ouverte ou fermée), maquillage virtuel, lecture labiale.

In recent years, the analysis of faces is a growing interest in the field of computer vision. The face is a powerful communications medium between human beings. It provides relevant clues on person identity, emotions or what it says. The GIPSA-lab carried out several studies on facial feature segmentation for multimedia applications (mixed reality, telephone terminal, human computer interaction, gesture interpretation for nonverbal communication interpretation, interactive driving simulator...). Studies deal with face location, with eye, eyebrow and mandibular arch contour extraction and, more recently, with mouth contour segmentation.

This work introduces an automatic outer and inner lip contour segmentation method for static images and video sequences. The algorithm is composed of two modules: a static module and a tracking module.

In case of static images, the first steps are face location and mouth bounding box extraction. Then, the lip contours are detected by combining active contours and parametric models. The jumping snakes are used to find key points to position a cubic curve parametric model which is appropriate to the possible lip shape deformations. Two inner parametric models have been built: one model for open mouths and another for closed mouths. Finally, maximization of relevant gradient flows is used to optimize the model parameter estimation.

In case of video sequences, the same static process is carried out on the first frame to initialize the tracking algorithm. On subsequent images, the tracking method is based on key point tracking techniques and the model is adjusted by the gradient flow maximization method.

The contributions of this work are: 1) a flexible lip parametric model, 2) several gradients combining luminance and chrominance information to highlight the lip contours, 3) quantitative and qualitative evaluation of the segmentation algorithm performances for the virtual make up and lipreading applications.

Keywords: Outer and inner lip contours, segmentation, tracking, active contours (jumping snakes), parametric models, cubic curves, tooth location, mouth state detection (open or closed), virtual make up, lipreading.