Segmentation des lèvres par un modèle déformable analytique

Nicolas Eveno

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité - Signal, Image, Parole, Télécoms

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

présentée et soutenue publiquement

par

Nicolas EVENO

le 14 novembre 2003

Segmentation des lèvres par un modèle déformable analytique

Directeur de thèse : Pierre-Yves COULON

JURY

M. James CROWLEY	Professeur	Président
M. Paul DELEGLISE	Professeur	Rapporteur
M. Jean-Luc DUGELAY	Professeur EURECOM	Rapporteur
M. Pierre-Yves COULON	Professeur	Directeur de thèse
Mme. Alice CAPLIER	Maître de Conférence	Co-encadrant
M. Gérard BAILLY	Directeur de Recherche	Examinateur

A Servane, à mes parents, à ma soeur...

Remerciements

Tout d'abord, je remercie mes encadrants Alice et Pierre-Yves pour la pertinence de leurs conseils, leur rigueur scientifique et leur soutien sans faille. Ils ont su me guider sans être trop directifs.
Si la rédaction d'une thèse n'est pas une sinécure, il en va de même pour sa lecture approfondie. Je remercie donc chaudement les membres du jury :

- James Crowley, pour avoir présidé la soutenance
- les rapporteurs, Jean-Luc Dugelay et Paul Deléglise, pour la finesse et l'exactitude de leurs analyses. Les questions qu'ils ont soulevées m'ont donné matière à réflexion pour encore quelques mois, ou quelques années...
- Gérard Bailly, pour les très nombreuses pistes qu'il m'a fournies.

Enfin, et surtout, je tiens à remercier Servane pour m'avoir supporté (dans tous les sens du terme) durant la rédaction de ce mémoire. Le lecteur devrait d'ailleurs également la remercier car, sans son expertise graphique, les pages qui suivent ne seraient probablement pas aussi claires.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les premiers mots échangés avec la machine furent d'abord inscrits sur des cartes perforées. Puis, avec l'arrivée des écrans et des claviers, le mode de communication se rapprocha de l'écriture et devint un peu plus humain. On commença même à parler de «langages» informatiques. L'évolution de ces langages fut très rapide et en quelques décennies on passa du très bas niveau (également appelé *langage machine*) à des niveaux beaucoup plus évolués et proches des structures syntaxiques et sémantiques humaines. Cette évolution fut dopée par la démocratisation du micro-ordinateur. Au début des années 50, une étude de marché restée célèbre évaluait le marché mondial à une cinquantaine de machines. Les utilisateurs étaient alors des experts capables de traduire leurs requêtes en langage de très bas niveau. Mais très rapidement, la réduction des coûts de fabrication permit d'envisager une distribution à beaucoup plus grande échelle. Se posa alors le problème de l'interface avec l'utilisateur non expert. Il s'agissait de masquer la très grande complexité de la machine derrière un certain nombre de couches logiciels et de créer finalement une *métaphore* simplifiée pour expliquer le fonctionnement et l'état du système ou pour provoquer de nouvelles actions.

La plupart des principes et des périphériques développés pour améliorer cette métaphore sont maintenant devenus très courants dans la conception des logiciels. Le principe le plus important a été le WYSIWYG (« What You See Is What You Get ») : l'image sur l'écran est toujours une représentation fidèle de la métaphore de l'utilisateur. Chaque manipulation de cette image entraîne une modification prévisible de l'état du système ou, tout au moins, tel que l'utilisateur l'imagine. Les éléments qui se sont imposés actuellement sont les fenêtres (*Windows*), les menus, les icônes et un outil de pointage. Les fenêtres rendent possible la représentation simultanée sur l'écran de plusieurs activités. Les menus permettent de choisir les prochaines actions. Les icônes représentent des objets informatiques sous forme concrète. L'outil de pointage, généralement la souris, permet de sélectionner les fenêtres, menus et icônes.

Aujourd'hui, plus de 500 millions d'ordinateurs sont installés dans le monde et, depuis 1995, il se vend plus de PC que de télévisions. La souris, le clavier ainsi que la représentation conceptuelle de l'état de l'ordinateur sont devenus naturels pour la plupart des utilisateurs. Cependant, bien que grandement facilitée, la communication avec la machine peut encore

paraître artificielle. Par exemple, les personnes âgées ou n'ayant que très peu de contacts avec les ordinateurs éprouvent souvent de grandes difficultés dans le maniement du clavier ou de la souris. Pire, pour les handicapés ces outils sont totalement inaccessibles. Ainsi, depuis quelques années de nombreuses recherches visent à faciliter l'accès à l'ordinateur, à le rendre plus *naturel*. Il ne s'agit plus pour l'homme de s'adapter à la machine, mais bien à la machine d'adopter les modes de communication humains. Dans *Les Robots*, Isaac Asimov écrivait à ce propos : «Les hommes ont fait la moitié du chemin. Les machines feront le reste»... Les premières avancées dans ce sens, dans les années 90, furent les logiciels grand public de dictée automatique ou de reconnaissance de caractères. Les puissances de calcul actuelles permettent d'envisager des analyses encore plus complexes portant désormais sur le corps humain lui-même. Etant donnés les efforts de recherche et les progrès récents dans ce domaine, on peut d'ores et déjà s'attendre à ce que les ordinateurs obéissent bientôt «au doigt et à l'oeil», au sens propre comme au figuré.

La zone du corps la plus chargée de sens est sans conteste le visage. En particulier, nous verrons, dans le premier chapitre de ce mémoire, qu'il est possible d'extraire une très grande quantité d'information de la zone de bouche. Nous montrerons que les applications de l'analyse faciale sont très nombreuses, et vont de la reconnaissance de la parole à l'identification du locuteur, en passant par l'interprétation des émotions et la création de clones synthétiques réalistes.

Dans le second chapitre, nous détaillerons les diverses techniques permettant d'extraire le contour des lèvres. Selon les types d'informations et de contraintes qu'elles utilisent, nous les classerons en trois grandes familles. Nous résumerons les forces et les faiblesses de chacune des méthodes. Partant de ces constatations, nous esquisserons les grandes lignes de notre algorithme.

La méthode de segmentation que nous proposons s'applique à des séquences vidéo et comporte 2 étapes principales : l'*initialisation* et le *suivi*. Dans le chapitre 3, nous traiterons de l'*initialisation*. Lors de cette phase, le contour externe des lèvres est extrait de la première image d'une séquence. Par conséquent, on parlera de «segmentation statique» car aucune information temporelle n'est utilisée.

Le quatrième chapitre détaillera l'étape de *suivi*, qui permet d'effectuer la segmentation dans les images suivantes. Durant cette étape, les résultats obtenus dans les images précédentes fournissent des informations supplémentaires susceptibles de rendre la segmentation plus robuste et plus rapide. Du fait de l'utilisation de ces informations temporelles, on parlera ici de «segmentation dynamique».

Enfin, dans le cinquième chapitre, nous résumerons les principales contributions de ce travail thèse et nous conclurons quant aux améliorations envisageables et aux perspectives pour la poursuite du projet.

CHAPITRE 1

Les applications de l'analyse labiale

1.1 Introduction

Depuis quelques années, on observe l'émergence d'une tendance générale visant à rendre plus naturels les rapports hommes-machines. Il revient désormais à la machine de comprendre le langage humain. Dès lors, l'intérêt de l'analyse faciale paraît évident. En effet, la zone du corps la plus chargée de sens est sans conteste le visage. La bouche produit la parole, la position des yeux renseigne sur l'objet ou la zone observés, les rides d'expression sont les miroirs de nos émotions... Bref, le visage est au centre des communications humaines. En particulier, de nombreuses études ont démontré qu'on pouvait extraire énormément d'informations de la bouche.

Tout d'abord, la bouche modèle le discours. Il existe donc un lien implicite très fort entre la parole *entendue* et la parole *vue*. Ce lien est d'ailleurs exploité très efficacement par les malentendants qui peuvent comprendre le discours par l'observation des mouvements des lèvres. La partie 1.2 détaille cet aspect et montre comment l'analyse labiale permet de concevoir des systèmes robustes de reconnaissance automatique de la parole.

Ensuite, à l'heure du virtuel, la visioconférence et les avatars de synthèse commencent à être courants. La recopie des mouvements de la bouche est nécessaire dans bien des systèmes de communication ou de compression. La partie 1.3 montre que la précision de l'analyse labiale prend une importance fondamentale lorsqu'il s'agit d'animer un clone de manière réaliste.

L'aspect *émotionnel* de la bouche est évident. Si les yeux peuvent également être rieurs ou tristes, on reconnaît en grande partie les émotions de nos interlocuteurs par la forme de leur bouche. Nous verrons dans la partie 1.4 que l'analyse labiale permettra peut-être de créer (entre autres !) des distributeurs de billets sensibles à nos états d'âmes.

Enfin, des recherches ont montré que la forme de la bouche ainsi que son évolution dynamique au cours du discours permettent d'effectuer de très bonnes identifications du locuteur. La partie 1.5 présente cette voie de recherche relativement récente.

1.2 Lèvres et communication

1.2.1 Aspect bimodal de la parole

Il a été montré que le cerveau effectuait une intégration des informations auditives et visuelles lors du processus de reconnaissance de la parole. L'effet McGurk illustre clairement ce phénomène [McGurk and McDonald, 1976]. Lorsqu'on présente des sources auditive et visuelle conflictuelles à un sujet, il arrive que le son perçu ne se rapproche ni de l'une ni de l'autre. Par exemple, lorsqu'une personne entend le son /ba/ et voit des lèvres former le son /ga/, elle aura l'impression d'entendre /da/. Quelques exemples de ce type sont réunis dans le tableau 1.1.

Les psychologues ont mené de nombreuses études sur l'effet McGurk. Ils ont montré qu'il existe également un effet McGurk «inverse», c'est-à-dire que la perception des mouvements de la bouche peut être altérée par la parole [Easton and Basala, 1982]. Au-delà des expériences artificielles décrites précédemment, cet effet a également été observé avec de la parole *naturelle*. De plus il se produit quelle que soit la langue utilisée et semble même influencer les perceptions des très jeunes enfants. Enfin, l'effet est robuste vis-à-vis de nombreuses conditions. Il continue à être observé dans le cas où il existe un décalage entre les stimuli, ou si le visage d'un homme est combiné avec la voix d'une femme. Pour plus de détails sur l'effet Mc Gurk, on peut se reporter à [AVSP, 1998]

Cet aspect bimodal de la perception de la parole se retrouve également dans la production. La parole est produite par la vibration des cordes vocales et par certains organes articulatoires tels que la trachée, la cavité nasale, les dents, le palais et les lèvres. Comme certains de ces organes sont visibles, il doit exister une relation implicite entre la parole *produite* et la parole *vue*.

L'unité acoustique élémentaire est le phonème. Dans [Rabiner and Juang, 1993], les auteurs dénombrent 48 phonèmes courants pour l'anglais. De la même manière, l'unité élémentaire visuelle est appelée le *visème*. Il existe de nombreux sons qui sont visuellement ambiguës. Ils sont donc associés aux mêmes visèmes. Par exemple, les phonèmes /p/, /b/ et /m/ sont tous produits par une bouche fermée et sont visuellement impossibles à distinguer. Ils constituent donc un seul et même visème. De la même manière, /f/ et /v/ sont visuellement identiques. Ils sont produits lors du contact des dents avec la lèvre inférieure. Ainsi, certains mots entiers seront très proches visuellement, comme «patte» et «batte».

La plupart des groupements de visèmes sont obtenus par l'analyse des confusions

Audio	Visuel	Perçu
ba	ga	da
pa	ga	ta
ma	ga	na

Tab. 1.1. quelques confusions dues à l'effet McGurk.

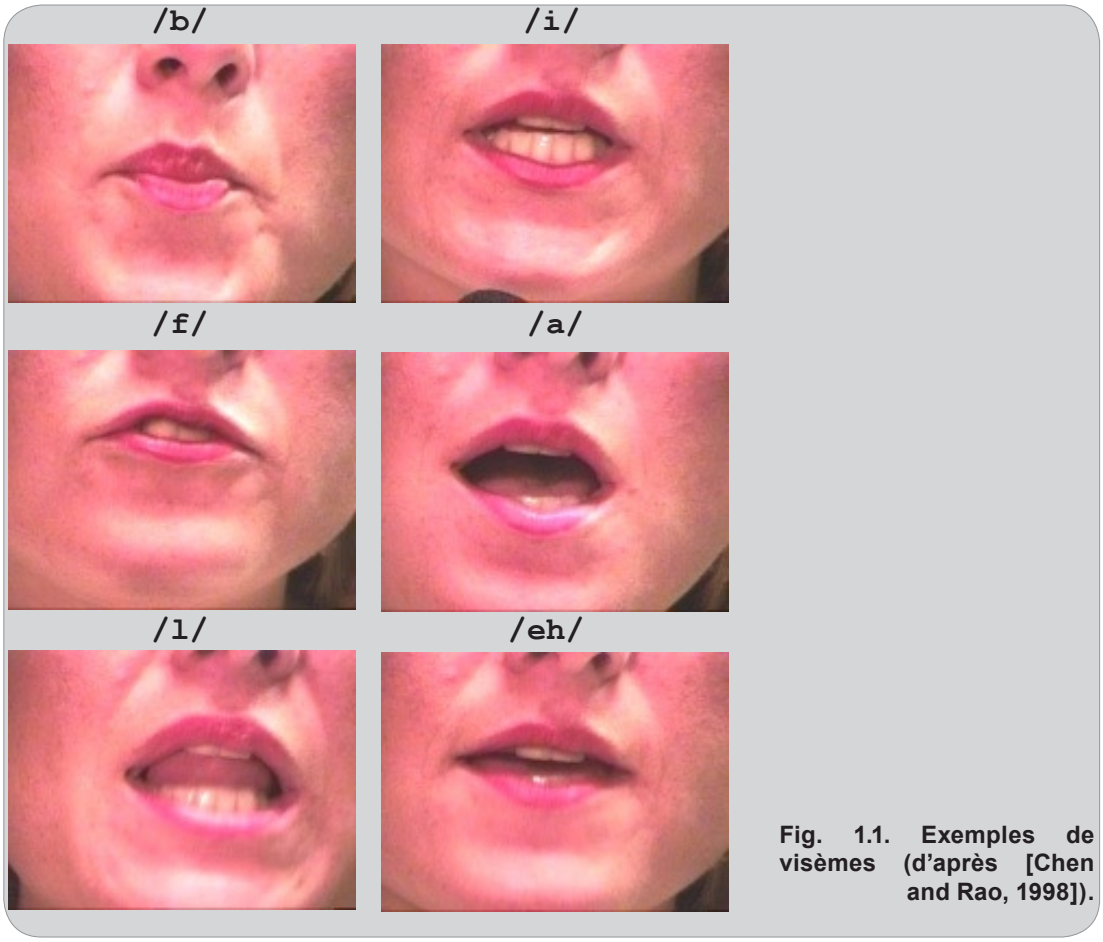


Fig. 1.1. Exemples de visèmes (d'après [Chen and Rao, 1998]).

observées dans les matrices de stimulus-réponse. Lors d'expériences, on demande aux sujets d'identifier visuellement des syllabes dans un contexte donné du type *VCV* (voyelle-consonne-voyelle). On génère ensuite les matrices de stimulus-réponse et on en extrait les groupes de visèmes. L'une des premières études des visèmes fut menée par Fisher [Fisher, 1968]. Il demanda d'identifier à la fois la consonne de début et la consonne de fin dans un contexte *CVC* (consonne-voyelle-consonne) et s'aperçut que les groupements de visèmes pour les consonnes initiales et finales étaient différents. En outre, il découvrit que la confusion entre les consonnes d'un même visème pouvait être directionnelle. Par exemple, le /n/ était souvent confondu avec le /t/, alors que le /t/ était rarement confondu avec le /n/.

Contrairement aux phonèmes, il n'existe pas actuellement de table standard des visèmes. Néanmoins, il est couramment admis que les visèmes des consonnes anglaises peuvent être regroupés en 9 familles distinctes [Dodd and Campbell, 1987], comme le montre le tableau 1.2. La figure 1.1 montre quelques-uns de ces visèmes. Les images de gauche sont associées à des consonnes, et celles de droite à des voyelles. En toute rigueur, on peut représenter certains visèmes de manière dynamique, c'est-à-dire sous la forme d'une séquence d'images. C'est sur-

1	**f, v**
2	**th, dh**
3	**s, z**
4	**sh, zh**
5	**p, b, m**
6	**w**
7	r
8	**g, k, n, t, d, y**
9	l

Tab. 1.2. Les 9 groupes de visèmes pour les consonnes anglaises.

tout vrai pour certaines voyelles comme, par exemple, le visème /ow/. Cependant, la plupart des visèmes peuvent être correctement approchés par des images fixes.

On distingue sans peine la majeure partie des voyelles, que ce soit visuellement ou auditivement. En revanche, ce n'est pas le cas pour les consonnes. Par exemple, les sons /b/ et /d/ sont très proches et il nous faut souvent recourir à «l'alphabet des noms» pour lever les ambiguïtés au téléphone («B» comme Bertrand, «D» comme Denise... etc). Il est d'ailleurs intéressant de noter que, dans ce cas précis, la confusion est très rare lors d'une conversation face à face. Le mouvement des lèvres permet de faire facilement la distinction entre les 2 sons. C'est encore une illustration de l'aspect bimodal de la parole. Cette observation a mené naturellement certains chercheurs à se demander si cette complémentarité des indices auditifs et visuels pouvait être utilisée pour rendre les systèmes de reconnaissance automatique de la parole plus robustes.

1.2.2 La reconnaissance automatique de la parole

Les premières tentatives d'intégration des indices visuels dans les systèmes de reconnaissance de la parole furent lancées pour résoudre les situations où l'audition n'était pas suffisante pour assurer la compréhension de la parole. Bien souvent, il s'agissait de rajouter quelques informations visuelles à l'intérieur d'un système de reconnaissance auditive permettant la détection de lettres séparées ou d'enchaînement *VCV* (voyelle-consonne-voyelle). La parole visuelle resta un certain temps une force d'appoint au système de perception auditif. L'échec des méthodes purement auditives pour la reconnaissance de la parole en environnement bruité, ainsi que l'accroissement des puissances de calcul, ont favorisé l'apparition de systèmes audiovisuels de reconnaissance automatique de la parole. Les recherches ont d'abord tenté de mettre en évidence l'apport de l'image pour la reconnaissance en environnement bruité. Comme le montre la figure 1.2, les taux de reconnaissance sont nettement supérieurs dans le cas où le signal audio est complété par quelques indices visuels. Une fois l'apport visuel clairement démontré, des études ont été menées sur l'amélioration de la reconnaissance de visèmes, d'enchaînement de lettres

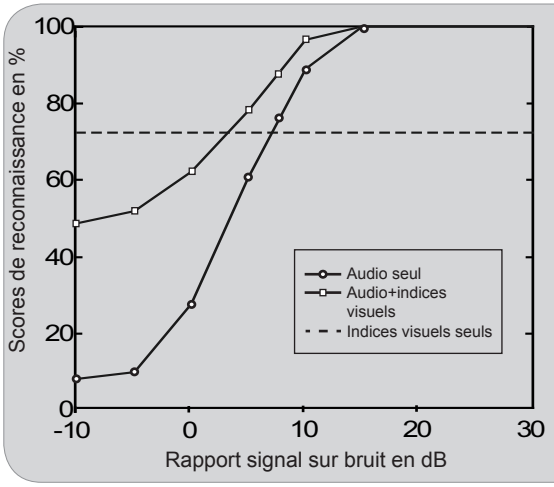


Fig. 1.2. Scores de reconnaissance d'un système automatique en présence de bruit (d'après [Chen and Rao, 1998]).

(*CVC*), de mots isolés et enfin de la parole continue [Chen and Rao, 1998]. La variabilité des méthodes employées est à la mesure de la variabilité des architectures possibles de reconnaissance. Il n'existe toujours pas de système de reconnaissance audiovisuel grand public alors que des études sont menées dans ce domaine depuis près de 20 ans. Cela dit, certaines étapes cruciales ont d'ores et déjà été franchies, et les bases méthodologiques commencent à être solides.

Typiquement, les systèmes de reconnaissance audiovisuels sont constitués d'une caméra, d'un ou plusieurs microphones et d'un système de traitement de la parole associé à un traitement de l'image. La fusion des données auditives et visuelles peut se faire ensuite de diverses façons. Les premiers systèmes étaient bâtis sur des architectures séquentielles, comme par exemple dans [Petajan, 1984]. Dans ce cas, l'audio est d'abord utilisé pour déterminer un certain nombre de candidats possibles ; puis l'image est utilisée pour lever les ambiguïtés. Plus tard, les systèmes ont évolué vers des architecture parallèles, où l'audio et le visuel sont analysés en même temps, la décision finale étant une fusion des deux résultats. Parmi les techniques de fusion de données, on distingue généralement deux grandes tendances. Les premiers systèmes de reconnaissance utilisaient des *modèles de Markov cachés*. Puis, certaines recherches ont montré que les *réseaux de neurones* pouvaient être également très efficaces. Selon les applications, on peut coupler la sortie de ce système à divers outils tels que des traitement de texte ou bien des systèmes de débruitage. Pour une comparaison des différent types de fusions possibles, on pourra se reporter à [Potamianos *et al.*, 2004].

Le premier système, proposé par Petajan dans les années 80 [Petajan, 1984], est basé sur l'analyse de la zone *intero-labiale* (située entre les lèvres). Après avoir détecté les trous de nez (qui sont, selon lui, les parties les plus facilement détectables du visage), la zone de bouche est obtenue par des mesures morphologiques. Un simple seuillage permet ensuite de faire ressortir les lèvres et d'estimer certains paramètres caractéristiques de la zone intero-labiale (hauteur, largeur, surface et périmètre). Ces paramètres sont ensuite utilisés dans un système multimodal séquentiel permettant la reconnaissance de lettres prononcées isolément. En utilisant les mêmes

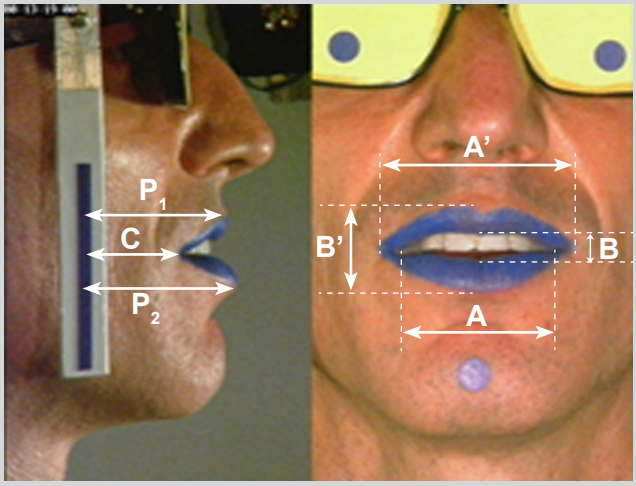


Fig. 1.3. Quelques-uns des paramètres utilisés par Lalouache. Les lèvres sont maquillées en bleu pour faciliter l'extraction (image *Institut de la Communication Parlée*).

indices visuels, Nishida réussit un peu plus tard à détecter le début et la fin de mots [Nishida, 1986]. Il fut le premier à remarquer que l'évolution dynamique des paramètres était susceptible de rendre la reconnaissance beaucoup plus robuste. Mase et Pentland arrivent d'ailleurs à la même conclusion en utilisant le *flux optique*, ou champ des vecteurs vitesse apparents [Mase and Pentland, 1991]. Dans [Goldshen, 1993], Goldshen utilise également cette propriété pour bâtir un système basé sur les modèles de Markov dont les entrées sont dominées par les dérivées temporelles des indices visuels. Dans le même esprit, les recherches menées par Stork *et al.* ont montré l'importance de la coarticulation pour la reconnaissance de la parole [Stork *et al.*, 1992].

Les indices visuels proposés par Petajan étaient à la mesure des puissances de calcul de l'époque et des outils disponibles en traitement d'image. Leur relative simplicité permettait un traitement proche du temps réel et rendait leur estimation aisée. Peu à peu, certain chercheurs ont tenté de confirmer leur pertinence ou d'améliorer la finesse de l'analyse visuelle. Pour palier aux éventuelles faiblesses du traitement d'image, Lallouache propose très tôt d'utiliser un maquillage bleu permettant une segmentation aisée des contours des lèvres [Lallouache, 1991] (voir figure 1.3). Les paramètres utilisés sont sensiblement les mêmes que ceux de Petajan, mais le maquillage permet de les obtenir avec une fiabilité nettement améliorée. Ce système permet d'ailleurs à Benoit d'identifier la vingtaine de visèmes du français [Benoît *et al., 1992*]. Dans [Finn and Montgomery, 1988], Finn et Montgomery utilisent des marqueurs autour de la bouche pour obtenir la position de certains points caractéristiques. Ils en extraient 14 distances caractéristiques dont l'analyse permet d'obtenir des taux de reconnaissance impressionnants sur des syllabes *VCV*. Il est à noter que, bien que ces méthodes paraissent artificielles du point de vue du traitement d'image, elles ont permis de démontrer rapidement la pertinence et l'utilité des informations visuelles.

Pour contourner la difficulté de l'extraction des paramètres visuels locaux, Yuhas *et al.* proposent un système très intéressant [Yuhas *et al.*, 1989]. Les valeurs de luminance des

pixels de la zone de bouche sont directement entrées dans un réseau de neurones chargé d'estimer le spectre audio correspondant. Une fusion de ce spectre avec le spectre mesuré permet ensuite d'effectuer une reconnaissance robuste. Ce système démontre que les informations visuelles peuvent être obtenues de manière plus globale. Potamianos a d'ailleurs proposé dans [Potamianos *et al.*, 1997] d'effectuer la reconnaissance en utilisant à la fois des indices locaux et des indices globaux (la transformée en ondelettes de la zone de bouche).

Une fois l'utilité des informations visuelles clairement démontrée, de nombreuses recherches ont été menées pour tenter de les obtenir automatiquement. L'accroissement des puissances de calcul amorcé dans les années 90 ainsi que l'apparition de nouvelles techniques de traitement d'image permettent d'envisager des algorithmes plus complexes et plus efficaces. Ainsi, les *contours actifs,* introduits par Kass et Witkin à la fin des années 80 [Kass *et al.*, 1987], ont rapidement séduit la communauté scientifique par leur formulation matricielle élégante et la possibilité qu'ils offrent de régler l'élasticité et la courbure des contours segmentés. Ils sont largement utilisés pour segmenter les lèvres, comme par exemple dans [Terzopoulos and Waters, 1993], [Leroy, 1996] ou [Delmas, 2000]. Les *modèles déformables* ont également un certain succès car en général ils permettent d'obtenir des formes plus réalistes que les contours actifs. Il existe de nombreux modèles déformables, allant du plus simple constitué de seulement 2 paraboles [Tian *et al.*, 2000], aux plus compliqués constitués de 4 ou 5 courbes polynomiales [Hennecke *et al.*, 1994] [Eveno *et al.*, 2002]. De même, les *formes actives*, basées sur un apprentissage des formes admissibles, peuvent être utilisées pour extraire des indices visuels [Luettin *et al.*, 1995]. Enfin, des systèmes d'analyse-synthèse basés sur un modèle 3D des lèvres ont également été proposés [Revéret *et al.*, 1997]. Nous présenterons plus en détails ces différentes approches dans le chapitre suivant.

Toutes ces techniques utilisant des informations de plus haut niveau ont permis une extraction relativement précise et fiable des paramètres géométriques de la bouche. A l'origine destinés aux systèmes de reconnaissance de la parole, les indices fournis ont rapidement été utilisés pour la création de visages parlants, comme le présente la partie suivante.

1.3 Synthèse de têtes parlantes

1.3.1 Motivations

Comme l'a montré la partie 1.2, l'utilisation d'indices visuels dans des systèmes de reconnaissance automatique de la parole permet une amélioration significative de la robustesse. Mais il ne faut pas oublier que ceci a d'abord été observé sur la compréhension *humaine* de la parole naturelle. Les malentendants, par exemple, utilisent très efficacement la forme et les mouvements des lèvres pour comprendre le discours. La conclusion de nombreuses études, comme celles présentées dans [Sumby and Pollack, 1954] ou [Erber, 1969], est que l'apport du visuel est d'autant plus significatif que l'audio est bruité ou défaillant. De plus, cette amélioration de la

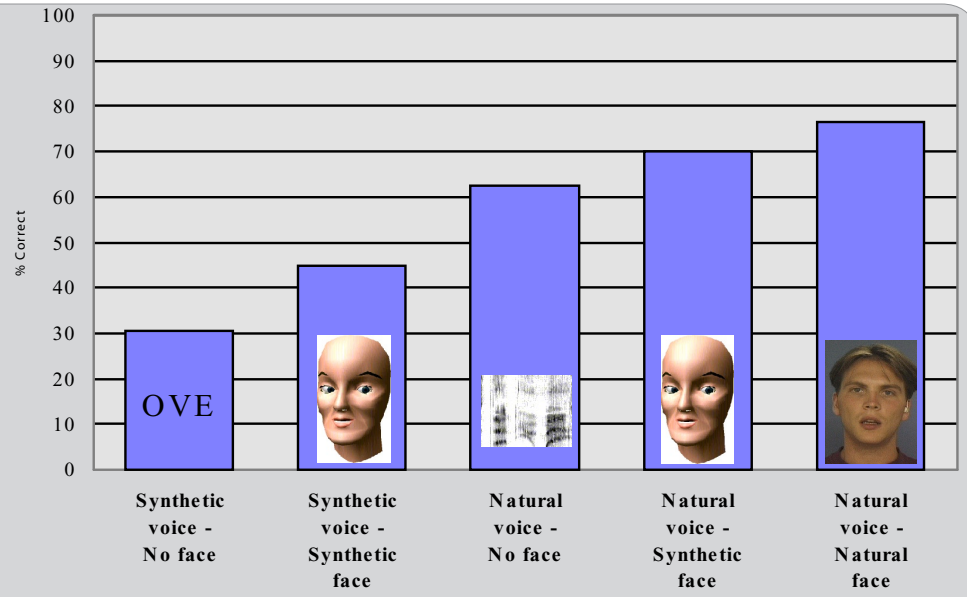


Fig. 1.4. Scores de reconnaissance par des sujets humains dans divers contextes (d'après [Beskow *et al.*, 1997]).

compréhension a également été observée en utilisant des images de têtes parlantes synthétiques [Cohen *et al.*, 1995]. La figure 1.4 montre les scores de reconnaissance moyens, obtenus dans [Beskow *et al.*, 1997], en environnement bruité (rapport signal sur bruit de 3dB). Il s'agissait pour des sujets humains de reconnaître des enchaînement VCV dans divers contextes : une voix synthétique seule, une voix et une tête synthétiques, une voix naturelle seule, une voix naturelle et une tête synthétique, et enfin une voix et une tête naturelles. Il est intéressant de noter que les scores obtenus avec une voix naturelle associée à une tête synthétique sont meilleurs que ceux obtenus par une voix naturelle seule. Dès lors, l'intérêt d'associer l'image à la parole paraît évident. En plus de permettre aux malentendants d'accéder à des outils de communication qui jusque là leur étaient interdits, l'utilisation de l'image pourrait améliorer la compréhension des sujet dits «normaux». Peut-être sonnera-elle le glas de «l'alphabet des noms» mentionné à la fin de la partie 1.2.1 !

De plus, la figure 1.4 montre également que le score de reconnaissance est meilleur avec une tête naturelle qu'avec une tête synthétique. Il semble que la compréhension soit d'autant meilleure que l'animation est réaliste. Pour obtenir des résultats optimaux, il faudrait donc transmettre directement l'image vidéo du locuteur, ce qui est très difficile techniquement. Même fortement compressée, la vidéo naturelle nécessite des bandes passantes qui sont, à l'heure actuelle, réservées aux grandes entreprises ou aux réseaux d'universités. La solution la plus réaliste est d'animer un avatar de synthèse, ou *clone*, en reproduisant le mieux possible les mouvements faciaux du locuteur. Ces derniers étant commandés par un nombre restreint de paramètres, la transmission peut se faire de manière fluide, même avec des connections relativement lentes. Le nouveau standard de compression vidéo MPEG-4 inclut d'ailleurs des

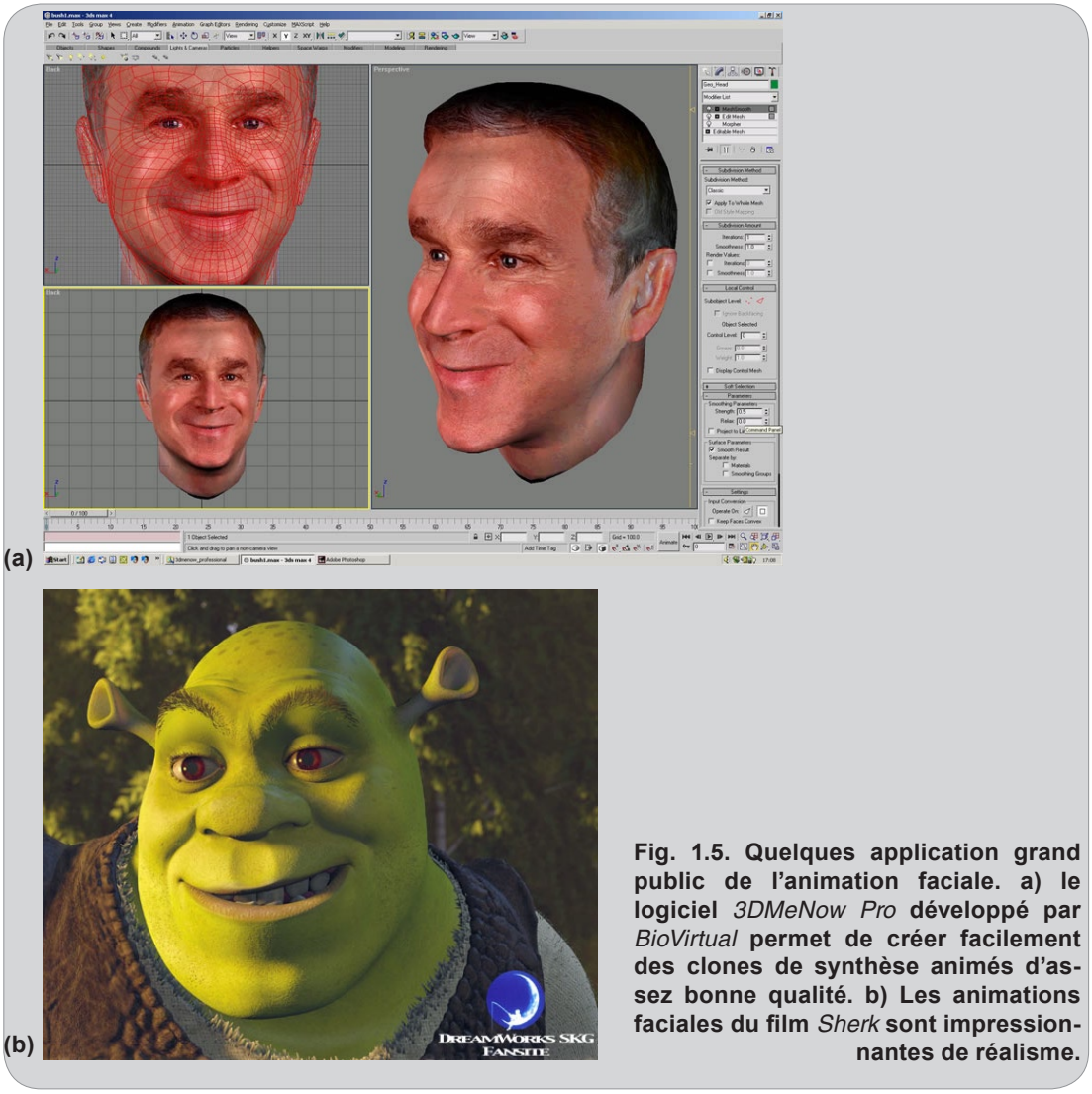


Fig. 1.5. Quelques application grand public de l'animation faciale. a) le logiciel *3DMeNow Pro* développé par *BioVirtual* permet de créer facilement des clones de synthèse animés d'assez bonne qualité. b) Les animations faciales du film *Sherk* sont impressionnantes de réalisme.

paramètres d'animation faciale destinés aux applications de vidéophonie. De même, des sociétés comme *Matrox* ou *BioVirtual* tentent de fournir au grand public des outils de synthèse et d'animation de clones, utilisables pour augmenter le réalisme des jeux en réseaux ou pour faire des présentations virtuelles à distance (voir figure 1.5-a). Dans tous ces systèmes, l'accent est mis sur l'animation de la zone de bouche car c'est souvent elle qui est la plus chargée de sens. Dans la partie 1.3.3, nous examinerons quelques-unes des stratégies adoptées pour effectuer des synthèses de clones.

Un autre domaine d'application possible de la synthèse de visages parlants est le cinéma. Ces dernières années, on a pu voir de nombreux personnages synthétiques héros de films d'animation et de publicités. Les premières synthèses grand public de qualité sont apparues au début des années 90 dans le film *Jurassic Park*. Les progrès des moteurs d'animation 3D ainsi

que l'accroissement des puissances de calcul ont permis de générer des images de dinosaures très réalistes. Si les mouvements du corps étaient correctement reproduits, il n'en était pas de même pour les visages. Comme les *visages* de dinosaures sont relativement simples et peu expressifs, cela ne posa pas de problème pour *Jurassic Park*. Il a fallu attendre quelques années pour atteindre un niveau de réalisme suffisant dans l'animation faciale. Dès que cela a été possible, de nombreux créateurs ont délaissé la pâte à modeler, les dessins animés ou les animations de figurines image par image pour se tourner vers des outils informatiques offrant une flexibilité et un degré de réalisme jamais vus jusque là. Aujourd'hui, des films comme *Shrek*, *Toys* ou *Monsters&Cie* mettent en scène des personnages aux mimiques faciales quasiment parfaites (voir figure 1.5-b). Les mouvements labiaux, en particulier, sont très bien synchronisés avec le discours et donnent réellement l'impression de se trouver face à un vrai acteur. Nous verrons un peu plus loin que, pour atteindre une telle perfection, les créateurs s'appuient sur des «béquilles» qui ne sont pas envisageables dans le domaine de la synthèse de clones grand public de visiophonie. De tels résultats ne pourront être reproduits *au naturel* qu'avec l'aide d'algorithmes d'analyse d'images très performants permettant de détecter avec précision les points caractéristiques du visage.

1.3.2 Le standard MPEG-4

A ce stade de notre dissertation, il convient de décrire plus en détails le format MPEG-4. Non pas parce qu'il s'agit d'un achèvement en terme d'animation faciale, mais plutôt parce qu'il promet d'être le premier standard grand public intégrant ce type de fonctionnalité.

Le groupe d'experts chargé de définir les formats de compression vidéo (*The Moving Pictures Expert Group - MPEG*) a récemment proposé un nouveau standard de compression vidéo d'un type totalement inédit : le MPEG-4 [MPEG, 1997]. Ce standard va au-delà de la compression *globale* dans le sens où il ne considère pas la séquence vidéo comme une succession d'images de taille constante associées à une bande son. Au lieu de cela, la séquence est vue comme un ensemble d'objets dynamiques spatialement et temporellement indépendants. Ainsi, ils peuvent être manipulés, stockés ou transférés de manière à composer une scène de toute pièce. Par exemple, il est possible de regarder les informations télévisées en ne gardant que le présentateur, le studio étant remplacé par une forêt équatoriale et la bande sonore rehaussée par des cris d'oiseaux exotiques. Il est même possible d'insérer dans un coin de son écran d'autres objets dynamiques tels que les cours de la bourse ou le classement en temps réel du tour de France ! Mais le MPEG-4 va encore plus loin puisqu'il intègre des possibilités d'animation de personnages. Grâce à un ensemble de paramètres normalisés, il est possible de décrire un grand nombre de postures corporelles ou faciales. Ainsi, il devrait même être possible de remplacer le présentateur du journal télévisé évoqué précédemment par un clone représentant le personnage de son choix (un singe, par exemple !). Dans ce cas, le terminal de l'utilisateur utilise simplement les paramètres d'animations extraits des mouvement du vrai présentateur et les plaque sur le clone. Au-delà de l'aspect ludique, ce standard de compression offre de réels avantages en terme de bande passante puisque, finalement, seule une quantité restreinte d'informations de

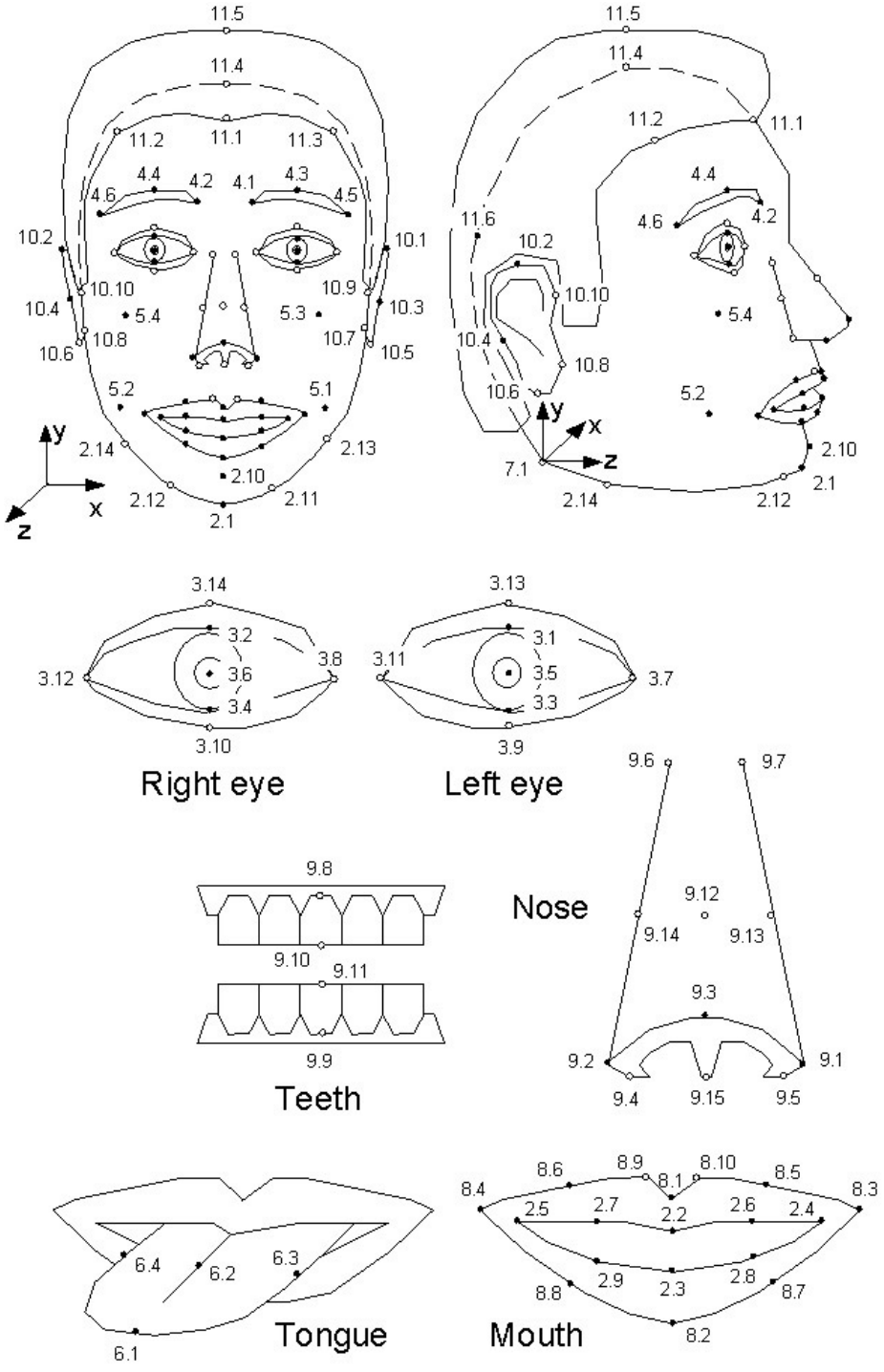


Fig. 1.6. Points caractéristiques utilisés pour décrire le visage dans MPEG-4, d'après [Delmas, 2000].

haut niveau est transmise.

Les paramètres utilisés autorisent une description très fine des postures, des tailles et des textures. Ils ont été choisis de manière à garantir la reproduction fidèle aussi bien des expressions «humainement possibles» que des expressions exagérées ou caricaturales (utilisées pour les personnages de dessins animés). Pour cela, trois types de paramètres ont été définis :

- **Les FP**, ou *Facial Definition Parameters*, sont un ensemble de points disposés sur le visage, comme indiqué à la figure 1.6. Les paramètres de positionnement de ces points sont notés ***FDP*** (*Facial Definition Parameters*). Ils permettent de personnaliser une animation, de la faire ressembler à tel ou tel personnage. Pour résumer, ils permettent de reproduire la topographie ou l'ossature d'un visage particulier sur laquelle sera plaquée la texture représentant la peau, les yeux, la barbe... En général, ils ne sont transmis qu'une seule fois par session, l'animation proprement dite étant ensuite assurée par les FAP.
- **Les FAP**, ou *Facial Animation Parameters*, assurent la reproduction des actions faciales de base. Ils contiennent en réalité 2 niveaux de description : les visèmes (21 pour le français) et les expressions (neutre, colère, dégoût, rire, tristesse, peur, surprise). Ils sont transmis sous la forme d'un flux continu de données permettant l'animation du modèle défini par les FDP.
- **Les FIT**, ou *Facial Interpolation Tables*, définissent la façon dont les FAP seront interpolés.

Bien que cette description puisse paraître complète, certains auteurs font remarquer que la déformation du modèle pose encore quelques problèmes. Dans [Bailly, 2001], Bailly souligne notamment le fait que l'extrapolation des positions de plusieurs dizaines de points à partir d'un seul pose encore problème. Il mentionne également le fait que l'animation réalisée ne tient pas compte de certains mouvements propres à l'élocution comme la protrusion des lèvres ou la rotation de la mâchoire. Cela a d'ailleurs conduit Vignoli et Braccini (dans [Vignoli and Braccini, 1999]) à proposer une couche supplémentaire de paramètres de contrôle susceptibles de combler le vide actuel de MPEG-4 : les *AP* (*Articulatory Parameters*).

1.3.3 Les différentes techniques

Comme expliqué en introduction de la section précédente, le MPEG-4 ne constitue pas l'état de l'art en terme d'animation faciale. Il existe plusieurs autres façons de créer un visage parlant, tout comme il existe plusieurs stratégies pour animer l'image.

1.3.3.1 La création de visages parlants

Il existe **2 voies principales** pour créer le visage proprement dit. La première approche s'appuie sur un maillage 3D du visage, la seconde sur une concaténation de morceaux de visages choisis dans une très grande base de données (un peu comme la composition d'un portrait robot dynamique). La **première famille** (à laquelle le MPEG-4 appartient) est issue des travaux de Parke. Au début des années 80, ce dernier proposa dans [Parke, 1982] de modéliser le visage à

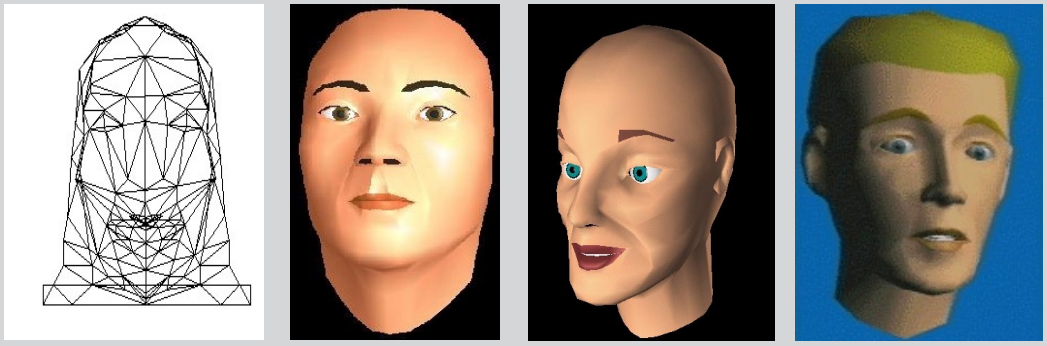


Fig 1.7. a) Le modèle «fil de fer» *Candide*. **On applique ensuite une texture sur le maillage pour obtenir une tête synthétique b)** *Baldi*, **c) la tête parlante du LCE, d)** *Sven*. **(D'après [Bailly, 2001]).**

l'aide d'un grand nombre de triangles. Au final, le visage ainsi fabriqué ressemble à un maillage serré de fil de fer (*wireframe model*) qui peut se déformer lorsqu'on déplace les sommets des triangles. La figure 1.7-a présente un modèle de cette famille, développé à l'université Linköping [Rydfalk, 1987]. En lui-même, le modèle fil de fer n'est qu'une représentation structurelle du visage. Pour le rendre plus naturel, il faut le recouvrir par une texture représentant la peau, les yeux et les cheveux. Cette texture peut être entièrement synthétique, comme pour les têtes des figures 1.7-b à 1.7-d. Elle peut également être produite par le placage d'une photo 2D sur le treillis 3D (*texture mapping*), un peu comme on emboutit une tôle plate sur un moule en relief. Pour des questions de réalisme et de ressemblance, c'est souvent cette dernière solution qui est privilégiée par les systèmes grand public de synthèse de clone. Par exemple, le clone de George Bush présenté à la figure 1.5-a a été réalisé de cette manière. Le visage de synthèse est ensuite animé en modifiant la position des points du maillage. Il est possible de ne contrôler qu'un seul point ou bien tout un ensemble de manière à synthétiser une expression ou un visème particulier.

Cette technique basée sur un maillage 3D *de surface* offre incontestablement une très grande flexibilité au visage. Comme chaque point peut être contrôlé individuellement, il est possible de déformer le visage à volonté. Or dans le cas de la synthèse d'un visage humain, cette très grande liberté peut être handicapante. En effet, l'espace de contrôle est gigantesque et une mauvaise combinaison de mouvements peut facilement mener à des expressions grotesques ou impossibles. Afin de rendre les animations plus réalistes, certains auteurs proposent de contrôler non pas les points de surface, mais plutôt les muscles et les organes responsables du mouvement. Le visage est alors modélisé par plusieurs couches en treillis ayant chacune leurs propres propriétés biomécaniques [Terzopoulos and Waters, 1990]. La figure 1.8 présente un exemple d'avatar de ce type. C'est d'ailleurs un modèle biomécanique très complet qui a permis aux

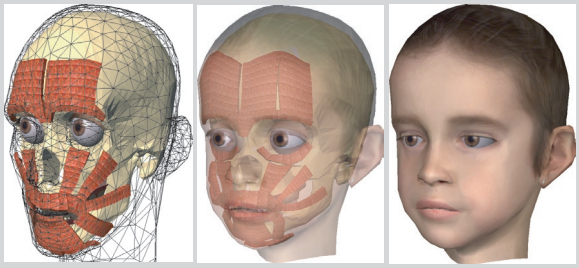


Fig 1.8. la modélisation biomécanique du visage donne des résultats très réalistes (d'après [Kähler *et al.*, 2002]).

studios *DreamWorks* de synthétiser *Shrek* de manière aussi réaliste. Sur la figure 1.5-b, on peut remarquer, par exemple, les plissements de la peau au-dessus des yeux, et les rides d'expression de chaque coté de la bouche. Ces artefacts ne peuvent être aussi bien reproduits que par la prise en compte des propriétés mécaniques de la peau.

Qu'ils soient biomécaniques en multi-couches ou bien de surface, les maillages 3D sont relativement difficiles à créer et à animer. La **deuxième approche** pour créer des visages parlants nécessite beaucoup moins de calculs au niveau du traitement d'image. Chaque image de l'animation est reconstituée en concaténant des morceaux de visages adéquats. Par exemple, les mouvements de la bouche sont reproduits en sélectionnant les bons visèmes dans une base de donnée. Les premiers systèmes de ce type donnaient des animations saccadées. De plus seule la zone de bouche était réellement animée, donnant des visages parlants assez peu réalistes où la bouche semblait «flotter» au milieu du visage. De nombreuses recherches ont tenté d'améliorer la fluidité en créant des images inter-visèmes par *image-warping* [Wolberg, 1990]. De même, les artefacts dus à la concaténation ont été effacés par des déformations et des lissages. Enfin, le nombre de parties mobiles a été augmenté pour obtenir un rendu plus réaliste des mouvements. Par exemple, le système développé par Cosatto et Graf utilise une décomposition en 6 parties [Cosatto and Graf, 1998] (voir figure 1.9). Au final, cette technique du «portrait-robot» basée sur une concaténation de morceaux de visages donne des résultats très réalistes et nécessite

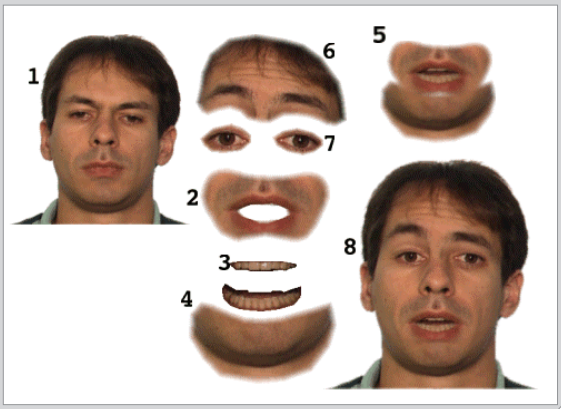


Fig 1.9. Le système proposé dans [Cosatto and Graf, 1998] repose sur une décomposition du visage en 6 zones. La concaténation dynamique de ces zones permet d'obtenir un visage parlant très réaliste.

assez peu de calculs. Cependant, pour la mettre en œuvre il faut disposer d'une base d'images relativement importante.

1.3.3.2 Les techniques de contrôle

Les techniques exposées précédemment permettent de créer le visage. Il faut ensuite l'animer, c'est-à-dire déformer le maillage (pour les visages orientés modèle) ou bien choisir les morceaux de visage adéquats (pour les techniques basées sur les échantillonnages faciaux).

Il est évident que, si on prononce /a/, la bouche est grande ouverte. De même, si on prononce /i/, elle s'étire vers l'extérieur. Partant de ces constatations, la plupart des recherches menées ont d'abord tenté d'utiliser les informations auditives pour déduire la forme de la bouche. Dans ce cas, le discours est analysé afin d'en extraire les mouvements probables du visage. Dans [Morishima *et al.*, 1989] par exemple, Morishima et ses collègues animent un modèle facial 3D en utilisant une *analyse cepstrale* du signal audio. La zone de bouche est commandée par 8 paramètres. Un entraînement du système est d'abord nécessaire pour lui apprendre à associer les cepstres aux mouvement labiaux. Cet exemple est assez représentatif des techniques utilisant l'audio seul, cependant il en existe de nombreuses variantes. Tout comme pour la reconnaissance automatique de la parole (voir section 1.2.2), on distingue 2 grandes familles d'analyse permettant d'associer le discours à des formes probables : celles qui utilisent les *modèles de Markov cachés* (comme dans [Welsh *et al.*, 1990]) et celles qui sont basées sur les *réseaux de neurones* (comme dans [Lavagetto, 1995]). Le principal intérêt des méthodes qui n'utilisent que les informations auditives est la possibilité de synthétiser une image lorsque seule la voix est disponible. Par exemple, cela peut permettre aux malentendants d'avoir un support visuel lors d'une conversation téléphonique [Beskow *et al.*, 1997].

Une autre possibilité pour animer un clone est d'utiliser directement l'image du locuteur lorsqu'elle est disponible. Bien que les motivations d'une telle démarche soient évidentes («animer une image à partir d'une image»), il a fallu attendre la démocratisation des systèmes d'acquisition et l'accroissement des puissances de calculs pour qu'elle produise des résultats acceptables. La plupart des méthodes d'analyse des mouvements des lèvres utilisées pour la reconnaissance automatique de la parole ont été reprises pour l'animation de clone. Ainsi, dans [Terzopoulos and Waters, 1993] Terzopoulos et Waters utilisent les *contours actifs* pour repérer et suivre quelques traits caractéristiques du visage. Ces derniers permettent ensuite d'animer un modèle 3D biomécanique. Dans [Essa and Pentland, 1994], le flux optique est utilisé pour déterminer les FAP. Dans [Botino, 2002], les yeux, les sourcils et la bouche sont suivis par les *modèles déformables* (voir figure 1.10). Des méthodes du type *analyse-synthèse* ont également été proposées. Dans ce cas, les paramètres d'animation sont trouvés en ajustant un modèle 3D à l'image acquise. Dans [Li et al., 1993], un modèle facial très simple (le modèle «Candide» de la figure 1.7-a) subit des déformations affines. Dans [Revéret *et al.*, 1997], c'est un modèle de lèvres qui est déformé pour reproduire les mouvements de la zone de bouche. Pour cela, les auteurs ont au préalable déterminé quelques axes principaux de déformation par *analyse en*

Fig 1.10. Détection de traits caractéristiques du visage par modèles déformables pour l'animation d'un clone (d'après [Botino, 2002]).

composantes principales. Dans [Odisio and Bailly, 2003], Odisio et Bailly proposent d'étendre à tout le visage la méthode utilisée dans [Revéret *et al*., 1997] pour suivre les lèvres.

La prise en compte de l'image permet de reproduire certains mouvements faciaux non (ou mal) détectés par le son, comme la direction du regard ou l'expression (joie, dégoût, peur...). Certes, l'analyse de l'intonation de la voix (ou *prosodie*) peut donner quelques indices quant à l'état d'esprit du locuteur et peut être utilisée pour rendre un avatar plus expressif. Cependant, des expressions complexes comme l'amertume ou le sourire ironique sont très mal rendues par ce type d'analyse. De même, les silences du locuteur peuvent être «gênés», «dubitatifs», «émerveillés»... sans que l'analyse audio soit capable de modifier correctement les traits du clone. Dès lors, non seulement la prise en compte de l'image permet de reproduire plus fidèlement les mouvements labiaux indispensables à la compréhension bimodale du discours, mais elle permet en plus de transmettre tout un panel d'émotions totalement ignorées par l'analyse audio. La partie suivante explique plus en détails comment l'analyse du visage, et en particulier de la zone des lèvres, peut permettre de reconnaître et de reproduire les émotions du locuteur.

1.4 Bouche et émotions

1.4.1 La communication paralinguistique

Les communications inter-personnelles utilisent 2 canaux distincts. Il existe tout d'abord un canal explicite (ou *linguistique*) utilisant la voix ou l'écriture et permettant la transmission du discours. En outre, il existe un second canal beaucoup plus complexe utilisant de nombreux supports comme le regard, la façon de bouger ou le ton de la voix. Ce canal est souvent qualifié de *paralinguistique* car il modifie, se substitue ou améliore la compréhension du discours parlé. Il permet d'obtenir des informations sur l'état d'esprit du locuteur ou d'accéder à un niveau plus subtil du discours. Un des supports privilégiés de la communication paralinguistique est le visage. Ainsi, aucun mot n'est nécessaire pour reconnaître la tristesse chez une personne dont les yeux et les coins de la bouche sont dirigés vers le bas. De même,

on reconnaît facilement l'ironie d'un locuteur qui affirme être bouleversé alors qu'il affiche un grand sourire.

De nombreuses études ont d'ores et déjà exploré le canal explicite, la principale voie d'investigation étant la reconnaissance automatique de la parole (décrite dans la partie 1.2). A l'opposé, le canal paralinguistique présente encore bien des zones d'ombre. Cependant, au-delà du besoin de rendre des clones plus expressifs, depuis quelques années on observe l'émergence d'une tendance générale visant à rendre plus naturels les rapports hommes-machines. Dans ce cadre, certaines recherches tentent de concevoir des systèmes capables d'analyser les émotions des utilisateurs. Par exemple, *Teradata*, une filiale de la compagnie japonaise NCR qui fabrique, entre autres, des terminaux de distribution de billets, s'est attelée à la tâche de rendre plus humains les distributeurs. En collaboration avec les chercheurs de l'université de Californie du Sud (USC) et dans le cadre du projet *E-Motions*, Teradata tente de produire un système capable de détecter quelques émotions de base chez l'usager. Ainsi, par des analyses de texture et par le repérage de quelques points particuliers situés notamment sur la bouche, les sourcils et le nez la machine devra pouvoir apporter des explications complémentaires aux personnes qui semblent angoissées. Elle pourra également accélérer le rythme d'affichage des instructions lorsque la personne semble impatiente ou agacée. La taille des caractères et le contraste pourront être augmentés lorsque l'usager plissera les yeux. Bref, l'objectif à terme est de créer un système qui ne réagisse plus seulement aux ordres tapés sur un clavier, mais directement, et de manière transparente, aux humeurs de l'utilisateur. L'analyse automatique des émotions peut également être utile aux psychiatres car de nombreux déséquilibres neurologiques peuvent être détectés sur le visage. Par exemple, la schizophrénie entraîne notamment une augmentation anormale du rythme des clignements des yeux. Au contraire la maladie de Parkinson se détecte très tôt par une baisse de ce rythme. De même, les émotions passagères apparaissant sur le visage d'une personne suivant une psychothérapie sont souvent très signifiantes pour le thérapeute. Ce dernier n'étant pas toujours attentif au bon moment, un système d'analyse d'émotions peut être très utile pour enregistrer les humeurs, même fugaces, du patient. Skip Rizzo, un psychologue de l'USC qui a travaillé sur le projet *E-Motions*, affirme que cela pourrait permettre de détecter un phénomène connu des psychiatre sous le nom de *visage du suicide*, apparaissant chez les personnes prêtes à se suicider et décrit par la littérature comme étant un visage anormalement neutre avec un regard «lointain».

1.4.2 Codification et reconnaissance des émotions

Comme l'a expliqué la partie précédente, depuis quelques années de nombreuses études tentent de concevoir des systèmes d'analyse automatique des émotions. Pour cela, il faut au préalable se donner un système de codification des mouvements faciaux relativement exhaustif. Dans [Ekman and Friesen, 1978], Ekman et Friesen ont proposé dès la fin ds années 70 un système de description appelé *FACS* (*Facial Action Coding System*). Avec ce système, les observateurs peuvent coder tous les mouvements faciaux possibles en combinant des mouvements

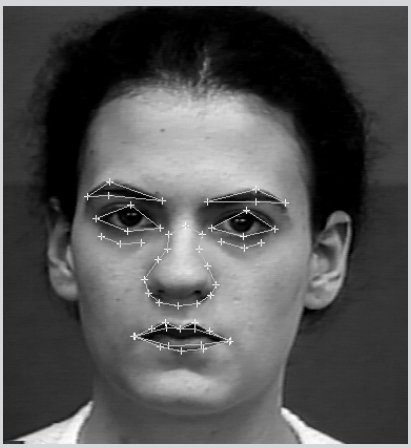


Fig. 1.11. Le suivi de points caractéristiques permet d'analyser les émotions du locuteur (d'après [Lien, 1998]).

élémentaires appelés *AU* (*Action Units*). Ekman et Friesen ont défini 44 AU, dont 30 sont reliés à la contraction de muscles spécifiques (comme, par exemple, le zygomatique) et 14 à des combinaisons plus complexes (comme la protrusion des lèvres). Il est important de noter que, bien que ses inventeurs proposèrent quelques combinaisons spécifiques pour représenter des expressions faciales classiques, le système *FACS* est purement descriptif et n'intègre aucun niveau d'interprétation des émotions. Ces dernières sont codées par des systèmes séparés comme *EMFACS* ou bien *MAX*. Grâce à leur exhaustivité et leur caractère descriptif, les *FACS* sont utilisés dans de très nombreux domaines comme l'animation (notamment par le standard MPEG-4), les neurosciences, la psychiatrie clinique et bien sûr l'analyse des émotions.

La plupart des recherches menées en reconnaissance d'émotions tentent de distinguer 5 ou 6 émotions de base (joie, peur, dégoût, colère, peur et parfois la surprise). Pour cela, différentes approches ont été proposées. Dans [Essa, 1995], un maillage 3D de visage est utilisé pour effectuer la reconnaissance. Son alignement avec le visage du locuteur permet d'estimer la contraction de 36 muscles faciaux correspondant à des *Action Units* du système *FACS*. A partir de cette estimation, un système de classification permet d'associer une des 5 émotions de base au visage observé. Dans [Mase and Pentland, 1991], Mase utilise le flux optique pour extraire les mouvements de 12 des 44 muscles faciaux. Les régions étudiées sont sélectionnées manuellement sur la première image, puis les directions des flux optiques moyens correspondant sont calculées dans les images suivantes. Dans [Yacoob and Davis, 1994], le flux optique est utilisé pour estimer directement les mouvements de régions entières, et non pas de muscles isolés. Les régions considérées (yeux, sourcils, nez et bouche) permettent d'effectuer une bonne discrimination des émotions. L'inconvénient des méthodes utilisant le flux optique est leur manque de sensibilité aux petits mouvements. Le moyennage sur des régions entières empêche très souvent la détection de contractions subtiles qui ont une grande importance pour l'interprétation des émotions. Pour estimer de manière plus sûre ces petits déplacements, certains auteurs préconisent d'utiliser le suivi de points. Au début des années 90, Himer *et al.* ont démontré la pertinence de cette approche en suivant des points caractéristiques dessinés sur le visage du

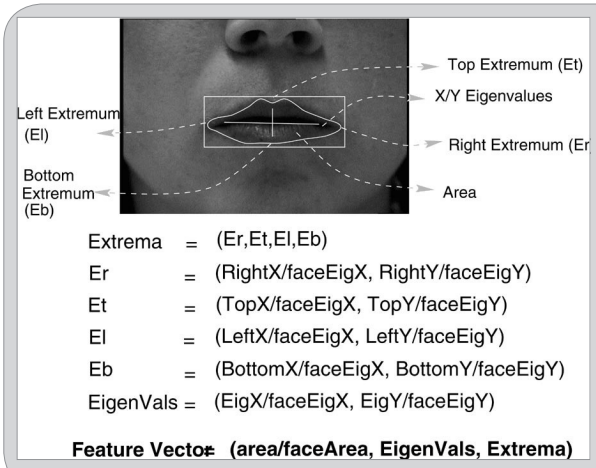


Fig. 1.12. Le système *LAFTER* extrait quelques paramètres géométriques des lèvres pour reconnaître les expressions (d'après [Oliver *et al.*, 2000]).

locuteur [Himer *et al.*, 1991]. Dans [Lien, 1998], Lien propose un système hybride basé sur le suivi automatique de points disposés autour du nez, de la bouche, des yeux et des sourcils (voir figure 1.11). Dans [Oliver *et al.*, 2000], Oliver, Pentland et Bérard n'utilisent que la forme de la bouche pour déterminer les émotions du locuteur. Pour cela, les lèvres sont segmentées et quelques paramètres sont extraits (voir figure 1.12). Ces paramètres sont ensuite utilisés pour effectuer une classification par modèles de Markov.

Il est intéressant de noter que, dans la majorité de ces méthodes, la zone de bouche joue un rôle prépondérant. Des études ont d'ailleurs montré que les parties les plus importantes pour la discrimination des émotions étaient les yeux et la bouche [Yamada, 1993][Morishima, 1995]. Dès lors, une analyse précise et robuste de la forme des lèvres est une étape nécessaire dans tout système évolué d'analyse automatique des émotions.

1.5 Bouches et identification

1.5.1 Généralités

Traditionnellement, les systèmes de contrôle d'accès nécessitent l'utilisation de cartes ou de mots de passe. Si les systèmes en eux-mêmes sont en général très fiables, il n'en est pas de même pour les utilisateurs. Les cartes peuvent être perdues ou volées, et les mots de passe peuvent s'oublier facilement. Pour minimiser les risques d'erreur, il faut utiliser des indices réellement caractéristiques de la personne qui ne puissent être ni perdus, ni volés, ni même altérés. Nous utilisons déjà de manière implicite certains de ces indices pour reconnaître au quotidien les personnes que nous côtoyons. Nous n'utilisons évidemment pas de mot de passe pour identifier un ami ! Le visage, la voix, la façon de bouger... sont autant d'indices efficaces que nous utilisons de manière instinctive.

Depuis quelques années, les systèmes de contrôle d'accès ont tendance à délaisser les

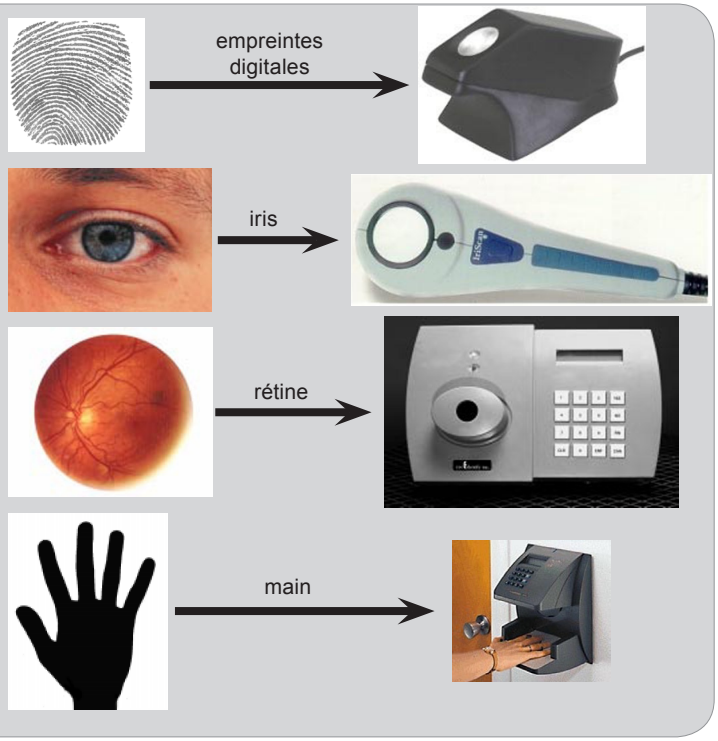


Fig. 1.13. La plupart des systèmes actuels utilisant des indices physiologiques nécessitent l'utilisation de périphériques spécifiques.

traditionnels *login* et *mots de passe* et tentent d'utiliser directement la biométrie des individus. Les grandeurs biométriques peuvent être définies comme des caractéristiques physiologiques ou comportementales utilisables pour vérifier l'identité d'une personne. Le but des techniques biométriques est d'automatiser la mesure de ces grandeurs. Selon les systèmes, plusieurs grandeurs peuvent être observées. Certaines sont physiologiques, comme les empreintes digitales, la rétine, l'iris, la géométrie de la main (voir figure 1.13), et d'autres sont comportementales, comme la voix ou la signature. Incontestablement, les mesures physiologiques sont beaucoup plus fiables que les mesures comportementales. Par exemple, d'après *International Biometric Group*, un cabinet de consultants en biométrie, l'analyse des empreintes digitales représente 34% des ventes de systèmes de vérification. Les systèmes basés sur la géométrie de la main représentent 26% du marché. Les autres indices physiologiques tels que l'iris ou la rétine sont un peu plus marginaux et sont en général réservés aux systèmes de haute sécurité. Au total, près de 80% des systèmes de reconnaissance sont basés sur l'analyse des indices physiologiques. Les résultats fournis sont en général bons et le prix des systèmes est en baisse. Certains fabricants informatiques, comme *Compaq* ou *Toshiba*, envisagent même de placer une petite matrice active à empreinte digitale sur leurs ordinateurs portables.

Cependant, le gros inconvénient des indices physiologiques utilisés actuellement est qu'ils nécessitent tous l'utilisation de périphériques spécifiques, comme le montre la figure 1.13. La lecture des caractéristiques de l'iris ou des empreintes digitales ne peut pas être effectuée avec de simples web-cams. Dès lors l'accession du grand public aux techniques biométriques

est plus limitée par les coûts d'installation que par les performances des systèmes. De plus, les techniques actuelles sont relativement contraignantes. Pour être identifié, il faut généralement se plier à des protocoles assez peu naturels, comme plaquer son œil devant un objectif ou bien appuyer son pouce sur une plaque sensible. Toutes ces raisons ont poussé de nombreux chercheurs à développer des systèmes plus naturels, moins invasifs et nécessitant moins d'équipement. Comme il a été fait remarquer plus haut, nous reconnaissons nos amis par leur visage et par leur voix. De plus la fiabilité de cette identification est bonne. Il doit donc être possible de copier ces mécanismes de reconnaissance (ou au moins de s'en inspirer) pour créer un système automatique basé sur le visage.

Actuellement, des sociétés comme *ZN-Vision* ou *Visionics* commercialisent des systèmes de reconnaissance faciale. L'identification est rapide et naturelle. Ces arguments ont d'ailleurs séduit les responsables du zoo de Hanovre qui ont remplacé très récemment les traditionnels tourniquets à ticket par des bornes équipées de caméras. Les visiteurs se présentent face à ces bornes, leur visage est comparé à ceux présents dans une base de données et l'accès est accordé ou refusé. Le processus est rapide et relativement fiable. De plus, les caméras installées dans les bornes d'entrée peuvent être également utilisées pour surveiller le zoo après la fermeture. D'autres systèmes utilisent la voix pour effectuer l'identification. Dans ce cas, une phrase particulière est prononcée par le locuteur face à un micro. Le système compare les paramètres du signal à ceux d'une base de données et valide l'identité. Du fait de leurs coûts d'installation très réduits, de grands espoirs ont été placés dans les techniques audio. Cependant, bien que ce domaine soit exploré depuis de nombreuses années (dès le début des années 50, dans son roman autobiographique «le premier cercle» Soljenitsyne décrit les recherches effectuées par le KGB pour reconnaître automatiquement la voix des dissidents au téléphone ou à la radio), les résultats obtenus sont toujours assez peu fiables. Dans le monde de l'industrie, l'analyse audio s'est fait rapidement dépasser par l'analyse faciale. D'après *International Biometric Group*, les ventes de systèmes de reconnaissance faciale ont déjà dépassé celles des système basés sur la voix, bien qu'ils soient arrivés récemment sur le marché de l'identification.

De manière à combiner les avantages des techniques audio et faciales, certains auteurs préconisent d'utiliser le mouvement des lèvres au cours du discours pour effectuer l'identification. Les premiers systèmes de ce type permettent déjà d'obtenir des résultats meilleurs que pour chacune des techniques prise séparément. De plus le matériel nécessaire (une web-cam et un micro) n'est pas dédié et est relativement courant, ce qui devrait permettre une diffusion très rapide auprès du grand public. La partie suivante présente plus en détails les motivations et les différentes approches dans ce domaine.

1.5.2 L'identification par les lèvres

Choisir les lèvres pour bâtir un système d'identification peut paraître étrange, voire artificiel. En effet, il n'est pas évident que nous utilisions cet indice pour reconnaître nos interlocuteurs. Cependant, dans [Etemad and Chellapa, 1997] les auteurs présentent des preuves

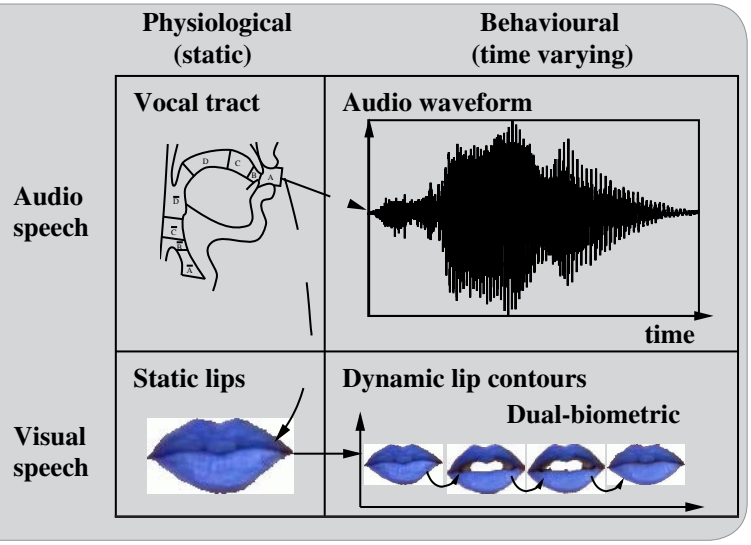


Fig. 1.14. Le système de Brand, présenté dans [Brand, 2001], utilise à la fois les informations audio et vidéo. La segmentation est facilitée par le *chroma key*.

expérimentales démontrant que la zone de bouche possède un pouvoir discriminant bien plus élevé que n'importe quelle autre partie du visage de même taille. Un autre avantage de l'analyse labiale est la possibilité qu'elle offre d'être combinée à des systèmes d'identification par la parole. Les parties précédentes ont déjà longuement exposé les liens étroits reliant la parole au mouvement des lèvres. En plus de faciliter la compréhension du discours, l'analyse bimodale du discours permet d'envisager des systèmes d'identification plus robustes que l'analyse audio seule. Dans sa thèse [Brand, 2001], Brand propose un système hybride, décrit schématiquement sur la figure 1.14, combinant l'analyse labiale et l'analyse audio. Il démontre notamment que les performances de l'identification bimodale sont supérieures à celles obtenues avec l'un ou l'autre des 2 systèmes pris séparément. De plus, il souligne le fait que l'analyse audio est facilement mise en défaut par des altérations passagères de la voix (comme celles causées par un rhume) ou bien par un environnement trop bruité, alors que l'analyse labiale reste neutre vis-à-vis de ces phénomènes. Enfin, un autre avantage de l'identification par les lèvres est qu'elle peut s'appuyer sur 25 ans de recherche en segmentation des lèvres. Les systèmes initialement développés pour la lecture labiale ou pour l'animation ont atteint un degré de précision et de robustesse suffisant pour envisager de les utiliser pour l'identification.

Luettin fut le premier à proposer un système d'identification basé sur la forme des lèvres. Dans [Luettin *et al.*, 1996+], il utilise les *formes actifs* pour effectuer la segmentation et extraire des paramètres géométriques des lèvres. Sans utiliser les informations auditives, il atteint des taux de reconnaissance supérieurs à 90%. Dans sa thèse [Chibelushi, 1997], Chibelushi utilise également des paramètres géométriques des lèvres et compare les performances de son système avec celles d'algorithmes purement audio. Il démontre que l'analyse vidéo des lèvres permet d'obtenir des scores au moins aussi élevés qu'avec l'analyse audio seule. Dans [Jourlin *et al.*, 1997], Jourlin *et al.* utilisent un système hybride combinant à la fois la forme des lèvres et l'information audio. Contrairement à Chibelushi, ils déduisent de leurs expériences que les

meilleurs scores d'identification sont obtenus en accordant un poids beaucoup plus important à l'audio qu'à l'image. L'apparente opposition de ces 2 conclusions est explicable par le fait que Chibelushi a extrait manuellement les paramètres géométriques des lèvres, alors que Jourlin a utilisé un algorithme de segmentation automatique. D'ailleurs, dans [Brand, 2001], Brand déduit de cette observation que la segmentation doit être très précise pour pouvoir réaliser une identification correcte. De manière à rendre son système indépendant vis-à-vis de la précision de la segmentation, il utilise un maquillage bleu des lèvres, connu dans le domaine de l'analyse labiale sous le nom de *chroma key* (voir figure 1.14). Bien que ce procédé soit artificiel et totalement impossible à mettre en œuvre auprès du grand public, il a déjà permis à de nombreuses recherches de s'affranchir du problème difficile de la segmentation. Par exemple, dans [Benoît *et al., 1992*] il a permis à Benoît d'identifier les visèmes du français. Il a également permis à Brand de démontrer quelques propriétés intéressantes dans le domaine de l'identification. Il confirme notamment les résultats obtenus par Chibelushi en démontrant que l'analyse du mouvement des lèvres est au moins aussi discriminante que l'analyse audio. Il démontre également l'importance de l'aspect dynamique, ce que Nishida avait déjà découvert dans le domaine de la compréhension bimodale du discours [Nishida, 1986]. Enfin, il admet le fait que ses résultats sont *asymptotiques* puisqu'ils sont fondés sur une segmentation parfaite. Cela l'amène d'ailleurs à souligner l'importance fondamentale de l'étape de segmentation, sans laquelle l'identification par les lèvres ne pourra progresser.

1.6 Conclusion

Le travail de thèse présenté dans ce mémoire s'inscrit dans un projet *RNRT* plus vaste : le projet *TempoValse* (*Terminal Expérimental MPEG-4 POrtable de Visiophonie et Animation Labiale ScalablE*), dont le but est de réaliser un terminal portable basé sur la normalisation MPEG-4 et supportant des applications multimedia telles que la visiophonie et la labiophonie. Le système doit pouvoir fonctionner en temps réel, sans maquillage et dans des conditions d'illuminations non maîtrisées. Par conséquent, l'extraction des paramètres géométriques des lèvres doit être **robuste** et **rapide**. Ces indices visuels sont ensuite utilisés pour animer un clone de synthèse. Au niveau matériel, le projet prévoit d'utiliser une micro-caméra montée sur un casque solidaire de la tête du locuteur. Les images traitées ont donc un cadrage fixe et couvrent une zone allant des narines au cou.

Initialement, ce travail de thèse était donc plutôt lié aux domaines de la communication et de la synthèse de têtes parlantes. Cependant, les systèmes utilisant l'analyse labiale sont nombreux, comme l'a montré ce chapitre. Nous avons donc tenté d'élargir le champ des applications possibles en ajoutant des fortes contraintes de **précision**, indispensable pour la reconnaissance d'émotion et l'identification par les lèvres. Le chapitre suivant fera le point sur les techniques permettant d'extraire les contours labiaux. A la fin de cet état de l'art, nous esquisserons les grandes lignes de notre algorithme, qui devra être aussi rapide, robuste et précis que possible.

CHAPITRE 2

Etat de l'art

2.1 Introduction

Comme nous l'avons vu au chapitre précédent, l'analyse labiale a de très nombreuses applications. Diverses techniques ont déja été proposées pour extraire des informations visuelles associées à la zone de bouche. Selon les types d'informations et de contraintes qu'elles utilisent, on peut les classer en 3 grandes familles.

Les *méthodes de bas niveau* n'utilisent que les informations des pixels de l'image. Elles supposent que les pixels des lèvres possèdent des caractéristiques homogènes et différentes de celles de la peau. La partie 2.2 montre comment, en partant de ce postulat, il est possible de segmenter les lèvres.

Les *méthodes de niveau moyen* exploitent également les informations de l'image, mais elles intègrent en plus des contraintes de régularité. Cela permet d'obtenir des contours plus lisses et de réduire l'influence du bruit. La partie 2.3 détaille une méthode de ce type : *les contours actifs*.

Enfin, les *méthodes de haut niveau* sont basées sur des modèles caractéristiques des lèvres, obtenus de manière heuristique ou statistique. Ainsi, la segmentation aboutit toujours à une forme admissible. Nous présenterons dans la partie 2.4 les principales techniques permetant d'intégrer une contrainte de forme au processus de segmentation.

Dans la partie 2.5, nous résumerons les avantages et inconvénients de chacune des méthodes. Partant de ces constatations, et compte tenu des objectifs que nous nous sommes fixés, nous esquisserons également les grandes lignes de notre algorithme.

2.2 Les méthodes de bas niveau

Les techniques de segmentation dites "de bas niveau" ("pixel based methods" en anglais) utilisent uniquement les informations présentes dans l'image. Elles supposent que les pixels de l'objet à segmenter possèdent des caractéristiques homogènes et différentes du fond. Ainsi, la segmentation peut être effectuée par l'identification et la séparation des classes *objet* et *fond*. Pour cela, différentes solutions ont été proposées. Certaines utilisent un simple seuillage d'une grandeur colorimétrique, alors que d'autres mettent en œuvre des techniques de classification plus évoluées.

2.2.1 Les espaces couleur

Un espace couleur est une convention permettant de visualiser, décrire et créer des couleurs. La toute première normalisation internationale est le système XYZ, établi en 1931 par la *Commission Internationale de l'éclairage* (*CIE*). Issu de mesures psycho-visuelles, cet espace englobe toutes les couleurs visibles et, par conséquent, contient tous les espaces qui ont été proposés par la suite. Ses 3 composantes ont des valeurs comprises entre 0 et 1. Y représente la luminance et (X,Z) donnent l'information de chrominance. De manière à obtenir un système indépendant de l'illumination, la *CIE* a proposé le système xyz, basé sur une normalisation de XYZ :

$$\begin{cases} x = \dfrac{X}{X+Y+Z} \\ y = \dfrac{Y}{X+Y+Z} \\ z = \dfrac{Z}{X+Y+Z} \end{cases} \qquad \text{(eq. 2.1)}$$

Bien que XYZ permette de décrire toutes les couleurs visibles, d'autres espaces ont été proposés. Dans les pages qui suivent, nous décrivons brièvement les plus courants en analyse faciale.

L'espace RGB - Il s'agit d'un système additif de couleurs, basé sur la théorie trichromatique. Cette dernière est très commode pour représenter les images sur un écran à tube cathodique (*tri-cathodique*). L'espace RGB, bien que mathématiquement rigoureux et techniquement efficace, ne tient pas compte des non linéarités du système visuel humain. De plus, il est fortement dépendant du contexte d'acquisition et de restitution. Dans cet espace, chaque couleur est décomposée en une combinaison de rouge, de vert et de bleu :

$$\boldsymbol{C} = R\boldsymbol{r} + G\boldsymbol{g} + B\boldsymbol{b} \qquad \text{(eq. 2.2)}$$

où $\boldsymbol{C}$ est la couleur à décrire, $(\boldsymbol{r},\boldsymbol{g},\boldsymbol{b})$ sont les vecteurs de base de l'espace, et (R,G,B) sont les coordonnées de la couleur. Bien que très utilisé, cet espace est relativement anti-intuitif et la synthèse d'une couleur particulière par un mélange des primaires est souvent difficile.

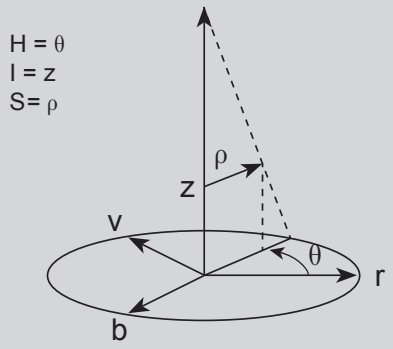


Fig. 2.1. Représentation du cône de couleur *HSI*.

Les systèmes TLS - Les espaces de ce type offrent une description des couleurs plus proche de la perception humaine. Parmi les nombreuses définitions existantes, on peut citer les espaces HSL (Hue Saturation Lightness), HSI (Hue Saturation Intensity), HSV (Hue Saturation Value), HCI (Hue Chroma Intensity) ... Bien que mathématiquement différents, tous ces espaces sont basés sur trois grandeurs colorimétriques facilement caractérisables :

- la teinte : attribut de la couleur (bleu, rouge, vert pomme ...)
- la saturation : le degré de différence entre la couleur observée et la couleur pure la plus proche (dont le spectre ne possède qu'une seule raie). Par exemple, on peut désaturer un bleu en lui ajoutant du blanc (dont le spectre contient toutes les fréquences). Le bleu pur devient ainsi pastel.
- la luminance : il s'agit de l'intensité de la radiation observée.

Classiquement, ces espaces sont représentés par un cône, comme le montre la figure 2.1. Les équations suivantes donnent un exemple de conversion en HSI :

$$\begin{cases} I = \dfrac{R+G+B}{3} \\ S = 1 - \dfrac{\min(R,G,B)}{I} \\ H = \dfrac{\pi}{2} - \arctan\left(\dfrac{2R-G-B}{\sqrt{3}(G-B)}\right) + k \end{cases} \qquad \text{(eq. 2.3)}$$

L'inconvénient principal de ces espaces est que la teinte (H) est définie par un angle. Cela rend le calcul coûteux (calcul trigonométrique nécessaire) et sensible au bruit. Malgré ces inconvénients, nous verrons dans la section suivante que la teinte reste une des grandeurs colorimétriques les plus utilisées pour la segmentation des lèvres.

YCrCb et formats vidéo - Les espaces suivants proviennent des normes standards en télévision analogique. Le format PAL (YUV) correspond à la norme européenne et le format NTSC (YIQ) à la norme américaine. Y est la luminance (tout comme dans l'espace XYZ) et (U,V) ou (I,Q) sont les composantes de chrominance. En général, on préfère le format YCrCb qui a l'avantage

de coder les chrominances de bleu (Cb) et de rouge (Cr) de façon explicite. Le principe de calcul repose sur la différence de la chrominance avec la luminance. Pour la recommandation de la CIE Rec.601-1, on obtient les relations de conversion suivantes :

$$\begin{bmatrix} Y \\ 1.4(Cr-128) \\ 1.8(Cb-128) \end{bmatrix} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad \text{(eq. 2.4)}$$

Il est courant de rencontrer des algorithmes de segmentation vidéo utilisant cet espace pour la simple raison que le flux de données est codé de cette façon.

Les espaces perceptifs - Le principal problème des espaces décrits précédemment est leur manque de linéarité. Une petite variation colorimétrique n'est pas perçue de la même façon sur tout le spectre. Pour résoudre ce problème, en 1976, la CIE définit deux nouveaux espaces couleurs plus proches de la perception humaine : Luv et Lab (aussi notés Lu*v* et La*b*, ou bien CIE-LUV et CIELAB). Ces espaces sont particulièrement adaptés aux systèmes de reproduction de haute qualité (reproduction de tableau, par exemple) et aux calculs précis sur des petites valeurs. Cependant, leur formulation non linéaire rend difficiles les conversions en temps réel. L'espace Luv est obtenu par les équations suivantes, où $X_nY_nZ_n$ représentent la référence de calcul (point blanc) :

$$\begin{cases} L = \begin{cases} 116(Y/Y_n)^{1/3} & si\ Y/Y_n > 0.008856 \\ 903.3Y/Y_n & si\ Y/Y_n \le 0.008856 \end{cases} \\ u = 13L(u' - u_n') \\ v = 13L(v' - v_n') \end{cases} \quad \text{(eq. 2.5)}$$

avec :

$$\begin{cases} u' = \dfrac{4X}{X+15Y+3Z} \\ v' = \dfrac{9Y}{X+15Y+3Z} \end{cases} \quad \text{(eq. 2.6)} \qquad \begin{cases} u_n' = \dfrac{4X_n}{X_n+15Y_n+3Z_n} \\ v_n' = \dfrac{9Y_n}{X_n+15Y_n+3Z_n} \end{cases} \quad \text{(eq. 2.7)}$$

L'espace Lab respecte les spécifications du format *TIFF* (*Tag Image File Format*). *a* et *b* sont respectivement les chrominances rouge/bleu et jaune/bleu. Il est également défini par rapport à XYZ et $X_nY_nZ_n$:

$$\begin{cases} L = \begin{cases} 116(Y/Y_n)^{1/3} - 16 & si\ Y/Y_n > 0.008856 \\ 903.3Y/Y_n & si\ Y/Y_n \le 0.008856 \end{cases} \\ f(t) = \begin{cases} t^{1/3} & si\ t > 0.008856 \\ 7.787t + 16/116 & si\ t \le 0.008856 \end{cases} \\ a = 500(f(X/X_n) - f(Y/Y_n)) \\ b = 200(f(Y/Y_n) - f(Z/Z_n)) \end{cases} \quad \text{(eq. 2.8)}$$

Fig. 2.2. Dans [Lyons *et al.*, 2003], un double seuillage sur les images issues d'une micro-caméra permet de segmenter la zone intero-labiale.

2.2.2 Le seuillage

La technique de segmentation *bas niveau* la plus évidente est le seuillage d'une grandeur colorimétrique. Le premier système de reconnaissance automatique de la parole, proposé par Petajan dans [Petajan, 1984], utilise d'ailleurs ce principe. Après avoir localisé les narines, la position de la zone de bouche est estimée par des mesures morphologiques. Un simple seuillage des niveaux de gris permet ensuite de faire ressortir les lèvres et d'estimer certains paramètres caractéristiques de la zone intero-labiale (hauteur, largeur, surface et périmètre). La même technique est mise en œuvre par Lyons dans [Lyons *et al.*, 2003] pour détecter l'ouverture de la bouche sur des images issues d'une micro-caméra montée sur un casque. Un double seuillage sur les niveaux de gris et sur la composante rouge permet de segmenter la zone intero-labiale (voir figure 2.2) puisque, selon Lyons, l'intérieur de la bouche est sombre et rouge. Une *analyse en composantes principales* sur les pixels obtenus permet ensuite d'extraire des paramètres de modulation utilisables par une interface musicale MIDI. Dans [Chiou and Hwang, 1997], Chiou et Hwang utilisent le quotient $Q=R/G$ pour caractériser les lèvres. Ils proposent d'effectuer un seuillage haut et bas de Q pour segmenter les lèvres. Les seuils sont déterminés expérimentalement et sont fixés une fois pour toutes, ce qui rend le système relativement dépendant du locuteur.

Si les conditions expérimentales sont bonnes, un simple seuillage permet donc de mesurer quelques paramètres géométriques des lèvres. Cependant, il permet difficilement d'extraire des contours fiables car le postulat utilisé (les lèvres sont caractérisées par une grandeur colorimétrique strictement bornée) est faux. De plus, la plupart des algorithmes de seuillage que nous avons rencontrés sont basés sur le système *RGB* qui est hautement dépendant des conditions d'illumination et des systèmes d'acquisition (voir section précédente). Ainsi, les seuils déterminés pour une base d'images ont peu de chances d'être réutilisables pour une autre base. Les seules segmentations de bonne qualité basées sur le seuillage résultent de l'utilisation d'un maquillage bleu (*chroma key*). Dans ce cas, les lèvres ont réellement des caractéristiques colorimétriques homogènes et très différentes de celles de la peau. Un simple seuillage sur la teinte permet alors d'obtenir des contours très précis. Bien que simpliste du point de vue du traitement d'image et peu réaliste au niveau pratique, cette méthode a permis d'analyser finement les mécanismes de production de la parole [Benoît *et al.*, 1992] ou bien de mettre en évidence le potentiel biométrique des lèvres [Chibelushi, 1997][Brand, 2001].

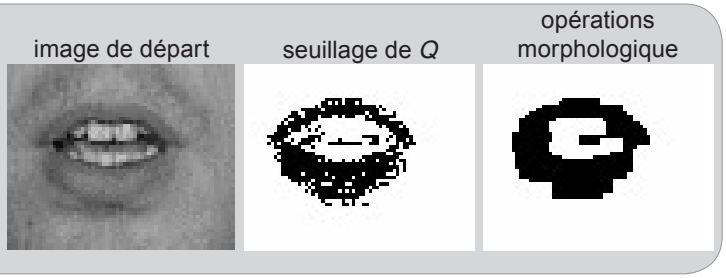


Fig. 2.3. Localisation des lèvres en utilisant Q, d'après [Wark and Sridharan, 1998].

Sans artifice, le seuillage ne permet pas de déterminer des contours fiables, mais il donne tout de même des indications sur les zones potentiellement intéressantes. Ainsi, il constitue souvent la première étape d'algorithmes plus complexes. Par exemple, dans [Wark and Sridharan, 1998], Wark *et al.* utilisent le quotient R/G pour localiser la zone de bouche (voir figure 2.3). Comme Chiou et Hwang, ils effectuent un seuillage haut et bas de Q. Mais ce seuillage est complété par des opérations morphologiques et par l'ajustement d'un modèle paramétrique sur la zone binaire extraite. Afin d'obtenir un masque binaire moins bruité et plus fiable, dans [Lucey *et al.*, 1999] la détection de la bouche est effectuée par un seuillage adaptatif sur Q. Dans ce cas, le seuil qui permet une séparation optimale des classes lèvres et peau est recalculé pour chaque image. La teinte est également très utilisée pour la localisation des lèvres. Dans [Zhang and Mersereau, 2000], Zhang et Mersereau comparent le pouvoir discriminant de plusieurs espaces couleurs. Ils remarquent que les composantes colorimétriques des lèvres et de la peau sont relativement uniformes dans les espaces (Cr,Cb) et (r,g,b). Cependant, leurs distributions se chevauchent beaucoup trop souvent, ce qui rend la segmentation difficile. A l'opposé, ils constatent que la teinte (H) des lèvres est relativement constante et bien séparée de celle de la peau. Ils proposent donc de localiser les lèvres par un seuillage sur la teinte. De plus, ils effectuent également un seuillage sur la saturation de manière à ne garder que les pixels fortement colorés. Dans une seconde étape, ils utilisent la pseudo teinte $R/(R+G)$ (introduite par Hulbert et Poggio dans [Hulbert and Poggio, 1998]) pour extraire les contours de l'image. Contrairement à la teinte classique, la pseudo teinte n'est pas une mesure angulaire et, de ce fait, son gradient est beaucoup moins bruité. Dans [Hsu *et al.*, 2002], Hsu *et al.* utilisent une transformation intéressante pour faire ressortir les lèvres. Ils montrent que $Cr/Cb\text{-}Cr^2$ est beaucoup plus important pour les lèvres que pour le reste du visage (voir figure 2.4), bien que l'homogénéité de cette expression nous paraisse douteuse. Après quelques opérations morphologiques, ils seuillent cette grandeur pour localiser la bouche.

2.2.3 La classification

De manière générale, la classification permet de séparer un ensemble en plusieurs groupes homogènes, ou *classes*. Il en existe 2 types : la *classification supervisée* et la *classification non supervisée*. La première, également appelée *analyse discriminante* en statistiques, nécessite d'avoir certaines connaissances a priori sur les classes. Pour cela, un entraînement

Fig. 2.4. Localisation des lèvres par le calcul de $Cr/Cb\text{-}Cr^2$, d'après [Hsu *et al.*, 2002].

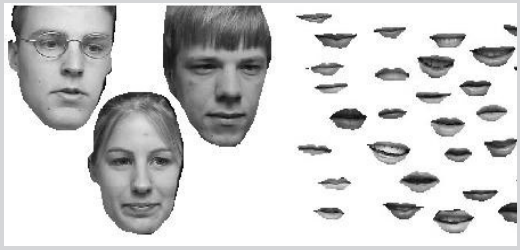
préalable du système est nécessaire pour dégager des règles caractéristiques des classes recherchées. Le deuxième type de classification, souvent appelée *cluster analysis* en anglais, permet de déterminer des classes sans aucun modèle a priori. Dans ce cas, l'algorithme tente de séparer l'ensemble qui lui est soumis en plusieurs classes homogènes. Ces techniques très générales ne sont pas réservées à l'analyse d'image et sont très variées. Les plus utilisées sont les *réseaux de neurones*, les *machines support vecteur*, les *analyses discriminantes linéaires* ou quadratiques, les *k-moyennes*, les *k-moyennes floues* ...

De manière à s'affranchir des problèmes liés au seuillage, certains auteurs proposent d'adopter une approche supervisée pour caractériser le mélange colorimétrique associé aux lèvres. Dans [Chan *et al.*, 1998], Chan *et al.* tentent de trouver la combinaison linéaire des composantes R, G et B permettant de séparer au mieux les lèvres et la peau. Pour cela, ils étiquettent manuellement des pixels de peau et de lèvres sur des images d'entraînement, et déterminent la combinaison optimale par une analyse statistique. L'image composite obtenue fait bien ressortir les lèvres et peut ensuite être utilisée pour effectuer la segmentation. Le même type d'approche est proposé par Nefian dans [Nefian *et al.*, 2002]. Il obtient la combinaison optimale par une *analyse discriminante linéaire* (*LDA*) et seuille l'image composite obtenue pour extraire la zone de lèvres. Dans [Vezhnevets, 2002], Vezhnevets effectue une caractérisation statistique de la couleur de la peau pour segmenter les lèvres. Lors de l'apprentissage, il divise l'image initiale d'une séquence vidéo en deux zones (*peau* et *non peau*) et calcule leurs histogrammes dans le plan (r,g). Il en déduit la grandeur *skin(r,g)* caractérisant le degré d'appartenance à la peau pour un pixel dont les composante r et g sont données. Enfin, il construit une image composite en calculant *lip* (équation 2.9) :

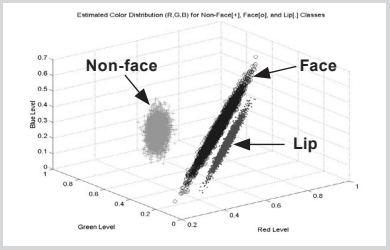
$$lip(x,y) = 0.07\frac{u}{v} + 0.3(1 - skin(r,g)) \qquad \text{(eq. 2.9)}$$

u et v sont les valeurs de la chrominance dans l'espace CIELUV, r et g sont les composantes rouge et verte du pixel (x,y), et les facteurs de pondération 0.07 et 0.3 sont déterminés empiriquement. Dans [Patterson *et al.*, 2003], les auteurs segmentent manuellement les lèvres dans quelques images de la base CUAVE. Ils en déduisent des approximations gaussiennes des distributions colorimétriques des classes *lèvres*, *visage* et *non visage*. Ces distributions de référence sont ensuite utilisées pour segmenter le visage et les lèvres grâce à un classifieur de Bayes (voir figure 2.5).

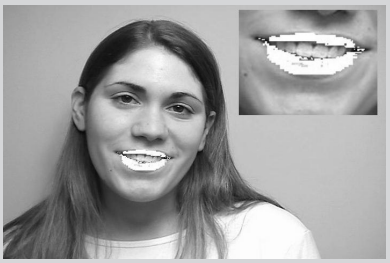
Les méthodes basées sur des classifications supervisées présentent quelques inconvé-



a)



b)



c)

Fig. 2.5. La segmentation manuelle (a) permet de calculer des approximations gaussiennes des distributions (b) qui sont ensuite utilisées pour reconnaître la classe *lèvre* (c), d'après [Patterson *et al.*, 2003].

nients. Tout d'abord, elles nécessitent un étiquetage manuel des images d'une base d'entraînement. Ensuite, les paramètres statistiques extraits sont dépendants des conditions d'illumination et d'acquisition. En effet, dans [Yang and Waibel, 1996], Yang et Waibel montrent que, même avec un éclairage constant, un changement de caméra peut produire des variations significatives des couleurs observées. Enfin, il est peu probable qu'une analyse statistique sur quelques dizaines de sujets différents permette de décrire toutes les couleurs de peau ou de lèvres possibles. Ainsi, il semble très difficile (impossible ?) d'établir un modèle statistique absolu permettant de reconnaître la "couleur lèvre" quelles que soient les conditions expérimentales.

Dans ces conditions, certains auteurs préconisent plutôt des approches *non supervisées*. Dans ce cas, aucune hypothèse sur les distributions statistiques a priori n'est faite. Même si nous n'avons pas tous la même couleur de lèvres, il existe toujours une différence entre la peau et les lèvres. Ainsi, il doit être possible de diviser le visage en deux régions relativement homogènes, même si les mélanges chromatiques qui leurs sont associés ne sont pas connus de manière absolue. Par exemple, dans [Liew *et al.*, 1999], Liew effectue une séparation de la peau et des lèvres par l'agrégation floue. A chaque pixel de l'image, il associe un vecteur chromatique caractéristique basé sur les espaces CIELUV et CIELAB. Ces espaces sont relativement uniformes (voir section 2.2.1) et beaucoup moins sensibles aux variations de luminosité que l'espace RGB. De plus, d'après les auteurs les distributions associées aux lèvres et à la peau sont bien séparées dans ces espaces. La figure 2.6-b présente la carte d'appartenance aux lèvres. Plus un pixel a des chances d'appartenir aux lèvres, plus il apparaît blanc sur cette carte. Un seuillage sur les degrés d'appartenance permet ensuite d'obtenir un masque binaire des lèvres (voir figure 2.6-c).

Les *champs de Markov aléatoires* (*MRF*) sont également utilisés pour effectuer des classifications non supervisées [Lievin and Luthon, 1999][Zhang and Mersereau, 2000]. Dans ce cas, le degré d'appartenance d'un pixel est influencé par ses voisins, ce qui permet de favoriser la compacité spatiale des classes. De plus, la théorie des champs de Markov permet d'inclure de

a) image de départ b) carte d'appartenance aux lèvres c) segmentation finale

Fig. 2.6. Segmentation des lèvres par *agrégation floue*, **d'après [Liew** *et al.*, **1999]**

nombreux indices en ajoutant simplement des termes à une fonction d'énergie. Ainsi, outre les interactions spatiales entre pixels, il est possible de combiner les composantes colorimétriques, les gradients [Zhang and Mersereau, 2000], le mouvement [Lievin and Luthon, 1999] ...

2.3 Une méthode de "niveau moyen" : les contours actifs

2.3.1 Cadre théorique

Les contours actifs (ou *snakes*) introduits par Kass et Witkin à la fin des années 80 [Kass *et al.*, 1987], ont rapidement séduit la communauté scientifique par leur formulation matricielle élégante et la possibilité qu'ils offrent de régler l'élasticité et la courbure des contours segmentés. Ce sont des courbes $\boldsymbol{\nu}$ définies paramétriquement ($\boldsymbol{\nu}(s)=(x(s),y(s))$, où s est l'abscisse curviligne) qui peuvent se déformer progressivement de manière à s'approcher au plus près des contours d'un objet. Cette déformation est guidée par la minimisation d'une fonctionnelle d'énergie comprenant 2 termes :

- une énergie intérieure E_{int} permettant de régulariser le contour ;
- une énergie externe E_{ext} reliée à l'image et aux contraintes particulières que l'on peut ajouter.

L'énergie interne dépend des dérivées du premier et du second ordre et est définie par :

$$E_{\mathbf{int}}(\boldsymbol{\nu}) = \frac{1}{2}\int_0^1 \alpha(s)\,|\boldsymbol{\nu}'(s)|^2+\beta(s)\,|\boldsymbol{\nu}''(s)|^2\,ds \qquad \text{(eq.2.10)}$$

Les termes en dérivée première de l'équation 2.10 ont pour effet de contrôler l'élasticité du contour, et les dérivées secondes permettent d'obtenir des courbes plus ou moins "lisses". Ainsi, on peut considérer le contour actif comme un élastique dont la tension et la courbure peuvent être réglées par les coefficients α et β. L'énergie externe du snake est déduite des caractéristiques de l'image et doit être minimale sur les contours. Pour une image en niveau de gris $I(x,y)$, une énergie externe très classique est définie comme suit :

$$E_{\mathbf{ext}}(x,y) = -|\nabla I(x,y)|^2 \qquad \text{(eq. 2.11)}$$

où ∇ est l'opérateur gradient. Le gradient peut également être précédé par un filtrage passe-bas de l'image. Cela permet d'obtenir des contours moins bruités et augmente leur zone d'influence. Ainsi, l'énergie totale du snake à minimiser est donnée par :

$$E_{tot}(\boldsymbol{\nu}) = E_{\mathbf{int}}(\boldsymbol{\nu}) + \int_0^1 E_{\mathbf{ext}}(\boldsymbol{\nu}(s))\, ds \qquad \text{(eq. 2.12)}$$

Minimiser cette fonctionnelle d'énergie revient à résoudre l'équation d'Euler suivante (en considérant les coefficients α et β constants) :

$$\alpha \boldsymbol{\nu}''(s) - \beta \boldsymbol{\nu}^{(4)}(s) - \nabla E_{ext} = 0 \qquad \text{(eq. 2.13)}$$

Cette équation peut être vue comme un équilibre de forces :

$$\mathbf{F}_{\mathbf{int}} + \mathbf{F}_{\mathbf{ext}} = 0 \qquad \text{(eq. 2.14)}$$

où $\mathbf{F}_{\mathbf{int}} = \alpha \boldsymbol{\nu}''(s) - \beta \boldsymbol{\nu}^{(4)}(s)$ et $\mathbf{F}_{\mathbf{ext}} = -\nabla E_{ext}$. La force interne empêche les élongations et les torsions trop importantes, et la force externe attire le snake vers les contours. En approximant les dérivées de l'équation 2.13 par des différences finies, puis en mettant les équations correspondantes sous forme matricielle, on obtient le schéma d'évolution suivant :

$$A\boldsymbol{\nu} - \mathbf{F}_{\mathbf{ext}}(\boldsymbol{\nu}) = 0 \qquad \text{(eq. 2.15)}$$

où A est une matrice à bande étroite dite pentadiagonale. Cette équation statique est résolue dynamiquement par l'ajout d'un terme de variation temporel nul à l'équilibre. L'équation 2.15 s'écrit alors :

$$A\boldsymbol{\nu}_i - \mathbf{F}_{\mathbf{ext}}(\boldsymbol{\nu}_{i-1}) = -\gamma(\boldsymbol{\nu}_i - \boldsymbol{\nu}_{i-1}) \qquad \text{(eq. 2.16)}$$

d'où :

$$\boldsymbol{\nu}_i = (A + \gamma I_d)^{-1}(\gamma \boldsymbol{\nu}_{i-1} + \mathbf{F}_{\mathbf{ext}}(\boldsymbol{\nu}_{i-1})) \qquad \text{(eq. 2.17)}$$

Le coefficient γ est souvent appelé coefficient d'amortissement et contrôle la vitesse de déplacement du snake. La position à l'itération *i* est alors déduite simplement de la force extérieure et de la position à l'itération *i-1*. Lorsque $\boldsymbol{\nu}_i$ et $\boldsymbol{\nu}_{i-1}$ sont très proches, on considère que la convergence est réalisée et le processus s'arrête.

2.3.2 Comportement

Les contours actifs sont un outil de segmentation puissant. Ils ont déjà été appliqués à de nombreux domaines comme le suivi de mouvement ou l'alignement d'images stéréo. Cependant, ils présentent 2 inconvénients majeurs. En premier lieu, les contours actifs sont «myopes», dans le sens où les forces qui les déforment ne dépendent que de leur très proche voisinage. Ainsi, si le contour est initialisé trop loin du contour final, il a peu de chances de le rejoindre. Ensuite, le réglage des coefficients n'est pas pris en charge théoriquement et est généralement effectué de façon heuristique.

Les paramètres d'élasticité et de courbure adaptés à une forme particulière ne sont donc

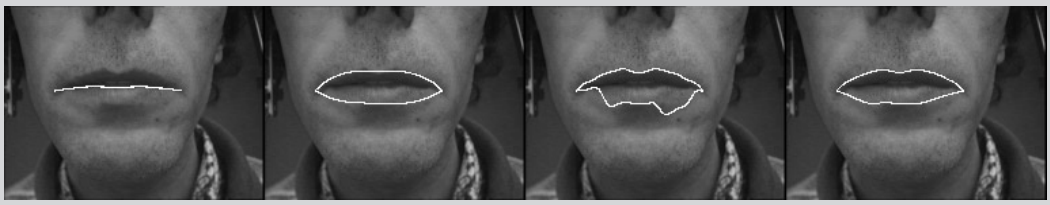
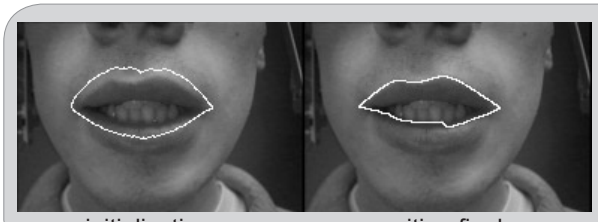


Fig. 2.7. Convergence d'un *snake* pour 4 valeurs différentes de α et β. De gauche à droite, les valeurs sont : (20;2),(2;20),(0.2;0.2),(2;2); d'après [Delmas, 2000].

pas utilisables sur une autre forme. La figure 2.7 illustre l'influence du choix des paramètres sur le contour final obtenu. On peut remarquer qu'une forte valeur de α induit des contraintes de tension importante et empêche le snake de suivre les contours souhaités. De la même manière, une forte valeur de β gomme les détails du contour. Il est donc nécessaire de procéder à des essais préalables pour déterminer le meilleur jeu de paramètres, ce qui est très handicapant dans le cas d'une segmentation automatique. Quelques rares auteurs ont tenté de s'attaquer au problème de la détermination automatique des paramètres. Par exemple, dans [Gao *et al*., 1998], Gao détermine des relations empiriques pour α et β en fonction de la distance entre 2 points successifs et de la courbure en chaque point du contour actif. Il montre que les résultats obtenus sont meilleurs grâce à cette adaptation des coefficients en cours de convergence. Néanmoins, il admet que sa technique est perfectible et qu'il y a encore du travail à faire dans cette voie.

La figure 2.8 montre l'effet d'une mauvaise initialisation sur le contour finalement obtenu. Comme il a été dit précédemment, le snake est *myope* car il est attiré par les contours proches de lui. De plus, si l'image est bruitée, il a de fortes chances de s'arrêter sur des contours parasites. De nombreux auteurs ont tenté d'améliorer la robustesse et la fiabilité des contours actifs [Caselles *et al*., 1995][Chakraborty *et al*., 1994][Cohen and Cohen, 1993][Leymarie and Levine, 1993]. Par exemple, Cohen [Cohen, 1991] propose d'utiliser une *force ballon* qui permet de *gonfler* ou de *dégonfler* le contour actif. Cette force permet au snake de dépasser les contours parasites et compense sa tendance naturelle à se contracter sur lui-même. Au final, la robustesse envers l'initialisation et le bruit est améliorée, mais une intervention humaine est encore nécessaire pour décider si le snake doit se gonfler ou se dégonfler. Amini *et al*. [Amini *et al.*, 1990] suggèrent d'utiliser la *programmation dynamique* pour minimiser la fonctionnelle d'énergie. Leur méthode recherche de manière exhaustive toutes les solutions admissibles, et



initialisation position finale

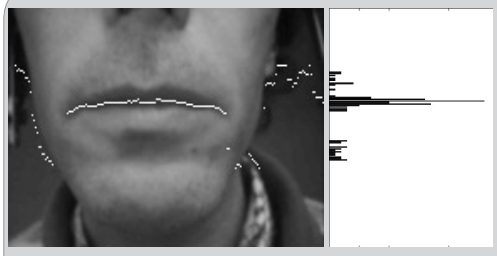
Fig. 2.8. Influence de l'initialisation sur le contour final, d'après [Delmas, 2000].

chaque itération conduit à un contour qui est un optimum local. Geiger *et al.* [Geiger *et al.*, 1993] proposent de résoudre le problème en une seule itération. Pour cela, ils autorisent la recherche du contour dans une zone très étendue autour de la position initiale. Neuenschwander *et al.* [Neuenschwander *et al.*, 1997] introduisent le *ziplock snake*, un nouveau type de contour actif qui se positionne progressivement à partir de ses deux points extrêmes (placés manuellement par l'utilisateur). Dans le même esprit, Berger et Mohr [Berger and Mohr, 1990] développent un snake qui grandit à partir d'un germe pouvant être ponctuel. Le contour est prolongé à chacune de ses extrémités par des segments de droite et est soumis à chaque fois à la minimisation de son énergie. Fua et Brechbuhler [Fua and Brechbuhler, 1996] définissent des contraintes fortes pour forcer le snake à passer par des points particuliers, ou tout du moins dans leur voisinage. Afin d'augmenter la zone d'influence des contours de l'image, Xu et Prince [Xu and Prince, 1998] introduisent une nouvelle force extérieure. Basée sur la diffusion isotropique des gradients de l'image, cette force permet également d'attirer le snake dans des régions jusque-là inaccessibles (formes «en creux», c'est-à-dire non convexes).

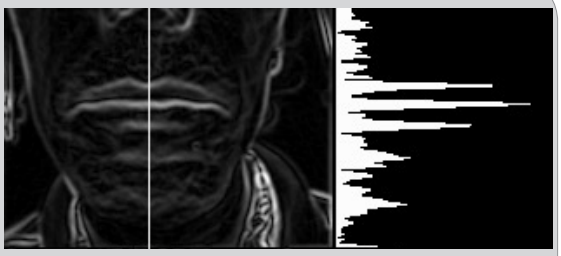
2.3.3 Application aux lèvres

La segmentation des lèvres par les contours actifs a été envisagée par de nombreux auteurs [Delmas, 2000][Leroy, 1996][Horbelt and Dugelay, 1995][Radeva and Marti, 1995]. Cependant, la tolologie particulière de la bouche ainsi que les contraintes de précision propres à l'analyse labiale ont poussé ces auteurs à proposer des adaptations spécifiques.

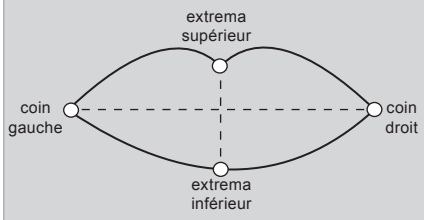
Comme nous l'avons déjà évoqué, l'initialisation du snake est sans doute le processus le plus important à résoudre dans une implantation optimale des contours actifs. Une initialisation trop éloignée des formes que l'on souhaite atteindre bloque le snake dans des zones de gradient parasite. De plus, la propension naturelle du snake à se contracter oblige à effectuer une initialisation très proche du contour recherché ou à l'extérieur de celui-ci. Initialisé trop loin, les coefficients de réglage du snake doivent être suffisamment grands pour permettre au contour actif de traverser les zones de gradients bruitées, mais suffisamment faibles pour ne pas sauter les zones de gradients recherchées. Diverses solutions ont été proposées pour effectuer l'initialisation proche des contours des lèvres. La plus classique est de localiser la bouche par projection des lignes et colonnes du plan intensité ou gradient de l'image, puis de rechercher des minima, des maxima ou des seuils sur les courbes obtenues. Dans sa thèse [Delmas, 2000], Patrice Delmas remarque que la zone interolabiale est toujours beaucoup plus sombre que le reste de l'image. Pour obtenir la position verticale de la bouche, il repère donc les minima d'intensité par colonne et en effectue une accumulation verticale (voir figure 2.9-a). De plus, il accorde un poids supérieur aux minima proches du centre car, dans son dispositif expérimental (casque fixé à la tête du locuteur), le visage du locuteur est au centre de l'image. Pour obtenir les commissures (limites horizontales de la bouche), il effectue un chaînage des minima par colonne en partant du centre de la bouche et détecte les premiers sauts importants de la chaîne. Ensuite, les limites haute et basse de la bouche sont obtenues par projection du gradient de luminance (voir



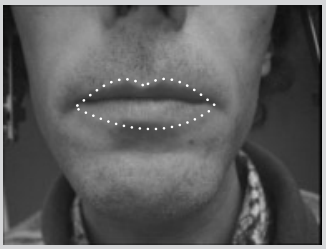
a) détermination de la position verticale de la bouche et des commissures.



b) détermination des limites haute et basse par la projection du gradient.



c) modèle utilisé pour l'initialisation du snake.



d) snake initial obtenu par échantillonnage du modèle.

Fig. 2.9. Algorithme d'initialisation du snake utilisé par Delmas [Delmas, 2000].

figure 2.9-b). Finalement, il adapte un modèle géométrique très simple (constitué de quartiques) sur les points limites qui ont été déterminés et l'échantillonne pour obtenir le snake initial (voir figure 2.9-c et 2.9-d). Radeva *et al.* [Radeva and Marti, 1995] adoptent une stratégie relativement proche de celle de Delmas. Leur localisation est également basée sur des projections des plans intensité. Cependant, ils ajoutent une contrainte de symétrie au visage pour rendre la détection plus robuste. Le snake initial est un modèle de lèvres positionné grossièrement et échantillonné. Très peu de détails sont donnés sur le type de modèle utilisé et sur son positionnement. Ces techniques de localisation basée sur des projections ont l'avantage d'être simples à comprendre et à mettre en œuvre. Cependant, Delmas affirme qu'elles sont efficaces seulement si les conditions expérimentales sont bien maîtrisées. Si l'éclairage, l'orientation, le cadrage ou le sujet changent, la localisation obtenue a peu de chances d'être correcte. Dans [Horbelt and Dugelay, 1995], Horbelt et Dugelay utilisent une méthode de localisation basée sur des méthodes de bas niveau de type seuillage, traitement morphologique et classification non supervisée (*clustering*). Bien que les auteurs ne mentionnent que "les ingrédients sans donner leur recette exacte", la localisation ainsi effectuée est probablement plus robuste que celle de Delmas ou Radeva car elle s'appuie sur des techniques éprouvées et relativement fiables. Un modèle déformable basique constitué d'ellipses est ensuite adapté aux lèvres. Son échantillonnage fournit les points du snake initial (voir figure 2.11).

Le second point délicat lié à l'utilisation des contours actifs pour la segmentation des lèvres est l'obtention d'un contour final qui suit correctement les contours de l'image et qui ressemble à une bouche. Comme les contours actifs sont des processus *à forme libre* (*free-form algorithm*), ils n'intègrent aucune connaissance a priori sur les formes admissibles. D'ailleurs, très peu d'auteurs se sont contentés d'appliquer les snakes aux lèvres sans ajouter quelques contraintes spécifiques permettant de guider leur évolution. Par exemple, un problème typique lié aux images de bouche, mentionné dans plusieurs études [Leroy, 1996][Petajan *et al.*, 1988][Radeva and Marti, 1995], est la détection précise des commissures. En effet il s'agit d'une zone à faible gradient sur laquelle les snakes ont beaucoup de mal à s'arrêter. Les forces extérieures y sont en général trop faibles pour compenser l'élasticité du contour actif. Si aucune précaution n'est prise, le contour final aura une forme "ronde" qui n'inclura pas les commissures. Pour résoudre ce problème, Delmas propose d'ajouter des forces ressorts empêchant le snake de trop s'en éloigner (voir figure 2.10). Il utilise pour cela l'estimation de la position des commissures obtenue par une technique exposée au paragraphe précédent. Bien qu'intéressante, cette méthode nécessite de définir une constante de raideur pour les ressorts utilisés. Une autre caractéristique des lèvres est leur courbure irrégulière. Certaines zones sont relativement "plates" (ou à faible courbure) alors que d'autres sont plus tortueuses (comme l'arc de Cupidon). Dès lors, le choix d'un jeu de paramètres ($\alpha;\beta$) constant sur tout le contour est un compromis entre la stabilité du snake et sa capacité à *coller* aux petits détails. La figure 2.7 illustre bien ce problème. Dans [Radeva and Marti, 1995] et [Radeva *et al.*, 1995], Radeva *et al.* proposent d'utiliser un modèle approximatif de la forme à segmenter pour diminuer l'influence du choix de α et β . Ils proposent pour cela une nouvelle définition de l'énergie interne :

$$E_{\text{int}}(\boldsymbol{\nu}) = \frac{1}{2}\int_0^1 \alpha(s)\,|\boldsymbol{\nu}'(s)-\boldsymbol{\nu}_0'(s)|^2+\beta(s)\,|\boldsymbol{\nu}''(s)-\boldsymbol{\nu}_0''(s)|^2\,ds \qquad \text{(eq. 2.18)}$$

où $\boldsymbol{\nu}_0$ est un ensemble de points issu de l'échantillonnage du modèle. Certes, cette méthode permet de diminuer l'influence des paramètres en imposant une référence de la dérivée et de la courbure en chaque point. Mais, même si elle est atténuée, l'influence du choix de α et β reste grande. De plus, cette technique est fortement dépendante de la qualité et de la précision du

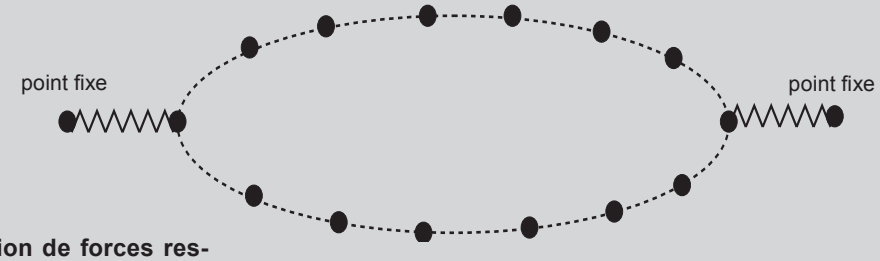


Fig. 2.10. Utilisation de forces ressorts pour attirer le snake près des commissures, d'après [Delmas, 2000].

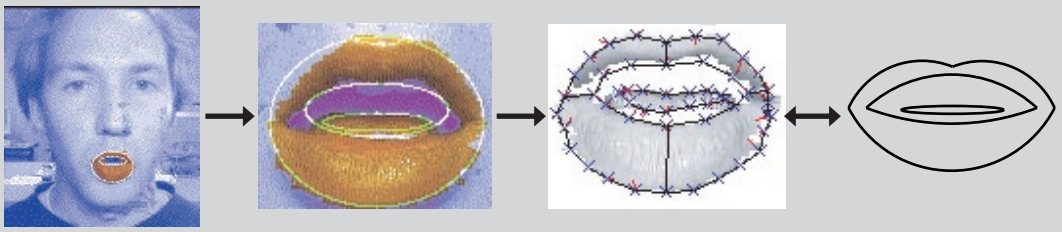


Fig. 2.11. Dans l'algorithme proposé par Horbelt [Horbelt and Dugelay, 1995], l'évolution du snake est guidée par un modèle de bouche.

modèle utilisé. La méthode proposée par Horbelt et Dugelay [Horbelt and Dugelay, 1995] est relativement proche dans le sens où la déformation du snake est également guidée par un modèle de lèvres. Leur algorithme (résumé à la figure 2.11) intègre des connaissances a priori sur la forme de la bouche, les contours actifs et les modèles déformables. L'aspect global des modèles déformables est associé aux capacités de localisation des contours actifs dans un processus bouclé permettant de tirer avantage des qualités de chaque méthode. Un modèle déformable, prenant en compte l'aspect étiré de la bouche, est appliqué sur une image de visage. Après convergence du modèle déformable, un contour actif est échantillonné sur la forme obtenue précédemment et est amené à évoluer afin d'obtenir des contours moins lisses. L'itération de ce processus, associée à des techniques de prédiction permet alors d'obtenir une détection fine des contours labiaux. Cependant, cette étude intéressante semble être restée au stade d'ébauche. Le choix des nombreux paramètres qu'elle nécessite n'est pas détaillé et aucun résultat quantitatif n'est donné. Afin de trouver le jeu de paramètres (α,β) optimal pour la forme des lèvres, Delmas effectue des tests d'apprentissage des coefficients sur images synthétiques, puis sur images réelles pour les lier aux valeurs du potentiel gradient vers lesquelles elles convergent. Son but est d'obtenir un arbre des valeurs optimales (en terme de rapidité et de qualité de convergence) de α et β en fonction du niveau moyen de gradient rencontré sur les contours recherchés afin de permettre un préréglage automatique de ces coefficients. Finalement, il obtient un couple de valeurs optimales très proche de celui qu'il avait déterminé auparavant de manière heuristique. Sa technique basée sur l'apprentissage ne lui permet donc pas d'éliminer les artefacts auxquels il avait été confronté jusque là. Il explique cet échec relatif par le fait qu'une approche rigoureuse devrait prendre en compte des valeurs de coefficients optimaux variables en tout point du snake. Toutefois, le nombre de variables à optimiser serait alors trop important pour espérer obtenir un quelconque résultat significatif.

2.4 Les méthodes de haut niveau

Les méthodes de moyen et de bas niveau décrites dans les sections précédentes sont des processus *à forme libre*. Elles n'intègrent aucune connaissance a priori sur les formes admissibles. A l'opposé, les méthodes de haut niveau sont basées sur des modèles caractéristiques des formes à segmenter, obtenus de manière heuristique ou statistique. Ces modèles génériques sont déformés de manière à être adaptés aux contours de l'objet. Les sections suivantes présentent 2 méthodes classiques en segmentation des lèvres.

2.4.1 Les modèles déformables analytiques

2.4.1.1 Principe

Introduits par Yuille au début des années 90 [Yuille *et al.*, 1992], les *modèles déformables analytiques* (*Deformable Templates*) permettent de décrire une forme de manière très compacte à l'aide d'un ensemble de courbes paramétrées. Par exemple, le modèle générique de lèvres proposé par Yuille (présenté à la figure 2.12) est constitué de 3 quartiques pour les contours extérieurs de la bouche et de 2 paraboles (confondues ou non, selon que la bouche est ouverte ou fermée) pour les contours internes. La variation des paramètres des courbes permet de faire évoluer le modèle correspondant vers les contours de l'objet à segmenter. Cette convergence est guidée par la minimisation d'une fonction d'énergie qui est, comme pour les *contours actifs*, la somme pondérée d'une énergie interne et d'une énergie externe. Dans le cas des *snakes*, l'énergie interne, paramétrée par les termes d'élasticité et de courbure, impose une contrainte relativement faible sur la forme finalement obtenue ("le contour doit être lisse et compact"). A l'opposé, l'énergie interne des *modèles déformables* permet de favoriser ou de pénaliser explicitement certaines déformations de la structure. Les contraintes exprimées par l'énergie interne seront donc fortement heuristiques. Ainsi, pour le modèle de bouche proposé par Yuille l'énergie interne favorise les comportements suivants :

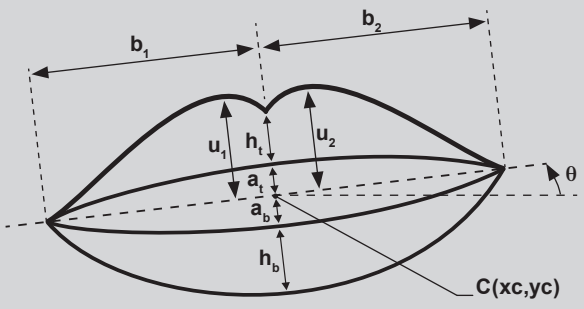


Fig. 2.12. Le modèle analytique de bouche utilisé par Yuille est composé de 3 quartiques extérieures et 2 paraboles intérieures.

- la symétrie du modèle de lèvre supérieure (2 quartiques supérieures similaires)
- le centrage de la bouche entre les 2 commissures
- un coefficient de proportionnalité constant entre les épaisseurs des lèvres supérieure et inférieure
- la cohésion du modèle en empêchant des mouvements verticaux trop importants de la lèvre supérieure pour éviter une trop grande proximité des régions bouche et nez.

L'énergie externe, ou d'adéquation aux données, attire le modèle vers les zones d'intérêt de l'image. Celle qui guide le modèle de Yuille utilise 3 champs énergétiques :

- le champ contour $\Psi_{contour}$ obtenu par la convolution d'un opérateur de lissage et d'un filtre gradient (Sobel, Prewitt, Canny...)
- le champ vallée $\Psi_{vallée}$ qui contient les zones sombres de l'image. Il est obtenu par seuillage bas du plan intensité
- le champ pic Ψ_{pic} qui comprend les zones claires de l'image. Il est obtenu par seuillage haut du plan intensité

Ces 3 champs correspondent aux principales caractéristiques visibles de l'image : les zones sombres, les zones claires et les zones de transition. Des considérations heuristiques permettent de les combiner pour obtenir l'expression de l'énergie externe. Par exemple, l'énergie externe associée aux contours externes des lèvres est surtout influencée par le champ contour car les zones correspondantes contiennent de fortes transitions. L'énergie liée aux contours intero-labiaux est calculée en utilisant le champ vallée, la transition entre les lèvres et l'intérieur de la bouche étant souvent sombre. Enfin, dans le cas d'une bouche ouverte, la détection de la présence des dents nécessite d'utiliser les champs contour et pic car les dents sont des zones claires bordées de forts gradients (transition entre les dents). Il est à noter que l'énergie associée au champ Ψ_{pic} est obtenue par une intégrale surfacique alors que les énergies *vallée* et *contour* sont issues d'intégrales curvilignes le long des courbes du modèle.

Une fois que les règles d'évolution du modèle ont été fixées (par l'intermédiaire des énergies internes et externes), il faut trouver les paramètres optimaux permettant de rapprocher les courbes des contours de l'image. Classiquement, cette minimisation est réalisée par un algorithme de *descente de gradient*. Le modèle de Yuille est régi par 10 paramètres : les coordonnées (x_c,y_c) du centre C, l'angle d'inclinaison θ, la largeur de la bouche b_1+b_2, les hauteurs des contours externes et internes en haut, h_t, et en bas, h_b, les paramètres d'aplatissement des quartiques supérieures u_1 et u_2. Une *descente de gradient* sur un tel nombre de paramètres peut être très longue et a toutes les chances de se "perdre" dans un minimum local. Aussi, Yuille propose d'adopter une méthode d'optimisation séquentielle bouclée de type "coarse to fine". Dans un premier temps, la position du modèle est optimisée : la bouche est centrée, orientée et sa largeur est calculée via l'optimisation des paramètres θ, x_c, y_c, b_1 et b_2. Puis, le positionnement plus fin des frontières est effectué par l'optimisation des 5 paramètres restants. Ce processus en 2 étapes est itéré jusqu'à la convergence du modèle.

La technique des *modèles déformables* proposée par Yuille a permis de palier aux faiblesses des *contours actifs* en introduisant la notion de modèle dans un processus de segmenta-

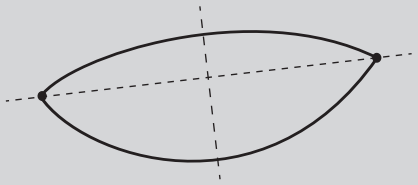
tion. Ainsi, le contour finalement obtenu fait forcément partie des formes admissibles. De plus, outre leur cohérence structurelle implicite, les modèles déformables sont peu sensibles au bruit de l'image car leur convergence est basée sur des calculs d'intégrales. L'influence d'une perturbation *locale* peut donc être compensée ou amoindrie par la minimisation de l'énergie *globale* sur la courbe correspondante. Cependant, malgré leur apport indéniable dans le domaine de la segmentation, les *modèles déformables* présentent quelques inconvénients. Tout d'abord, comme les *snakes*, ils sont relativement sensibles à la position initiale du modèle. S'il se trouve trop loin du contour final, la descente de gradient risque de s'arrêter dans un minimum local. Ensuite, le temps de convergence est en général très long car le nombre de paramètres à régler est important. Même avec l'optimisation séquentielle des paramètres, Yuille ne parvient pas à faire converger son modèle en moins de 5 minutes. Enfin, un inconvénient très souvent relevé dans la littérature est la trop grande rigidité du modèle. Les contraintes de symétrie l'empêchent de reproduire fidèlement une grande partie des formes générées lors du discours naturel. De même, les courbes utilisées ne sont généralement pas suffisamment déformables pour obtenir un contour "réaliste". La plupart du temps, elles permettent d'estimer les grandeurs caractéristiques classiques de l'analyse labiale (ouverture de la bouche, largeur,...), mais leur rigidité les empêche de suivre le contour des lèvres avec précision.

Les études contemporaines dans le domaine des modèles déformables concentrent donc principalement leurs efforts dans la résolution de 2 problèmes : l'accélération de l'algorithme et l'amélioration de la précision. Les modifications les plus significatives concernent le choix du modèle et la technique de déformation. Ces 2 points sont détaillés dans les sections suivantes.

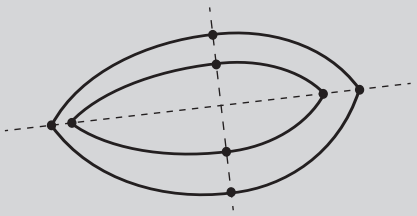
2.4.1.2 Les modèles utilisés

Le choix d'un modèle est un compromis entre la complexité algorithmique et la déformabilité. Pour obtenir un grand nombre de degrés de liberté (et donc un choix plus vaste de formes possibles), il faut soit utiliser un grand nombre de courbes, soit utiliser des courbes polynomiales d'ordre élevé. Les 2 solutions entraînent une augmentation du nombre de paramètres qui risque de rendre lente et difficile la minimisation de l'énergie. A l'opposé, la convergence sera très rapide si le modèle n'est composé que de 2 ou 3 courbes simples. Mais dans ce cas, la rigidité du modèle rendra la segmentation approximative.

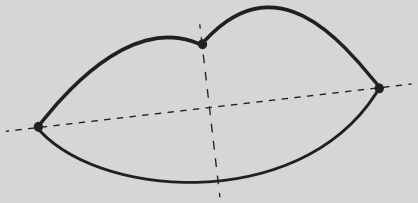
Un grand nombre de modèles de bouche a déjà été proposé dans la littérature. Par exemple, dans [Botino, 2002][Tian *et al.*, 2000] et [Rao and Mersereau, 1995] le contour extérieur des lèvres est modélisé très simplement par 2 paraboles (voir figure 2.13-a). Dans [Zhang, 1997], ce modèle est complété par une ou deux paraboles (selon que la bouche est ouverte ou fermée) pour représenter le contour intérieur. Le nombre réduit de paramètres de contrôle permet d'effectuer une segmentation rapide et simple. Cependant, dans [Botino, 2002] et [Tian *et al.*, 2000], les auteurs admettent que leur modèle est trop rigide et ne permet pas de reproduire avec précision les formes asymétriques notamment (voir figure 2.14). Pour rendre le modèle un peu plus flexible, Pantic *et al.* [Pantic *et al.*, 2001] proposent d'utiliser 4 paraboles pour l'exté-



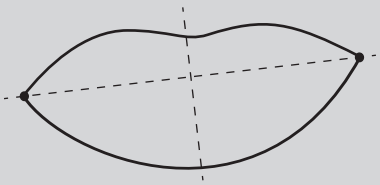
a) modèle symétrique à 2 paraboles, utilisé dans [Botino, 2002][Tian *et al*., 2000][Rao and Mersereau, 1995].



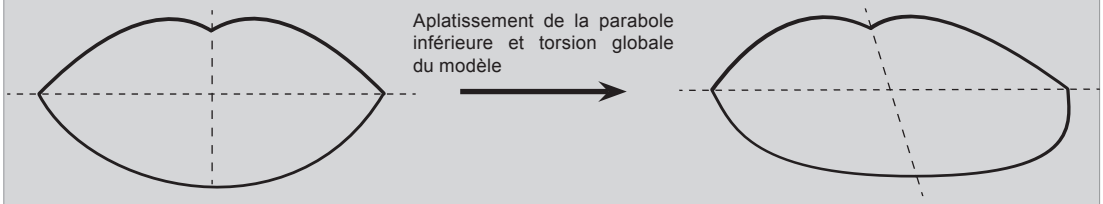
b) modèle à 8 paraboles, utilisé dans [Pantic *et al*., 2001].



c) modèle à 3 paraboles, utilisé dans [Coianiz *et al*., 1995].



d) modèle à 1 quartique et 1 parabole, utilisé dans [Wark and Sridharan, 1998]..



e) modèle symétrique à 3 paraboles, aplati et tordu, utilisé dans [Liew *et al*., 2000].

Fig. 2.13. Quelques exemples de modèles génériques de lèvres proposés dans la littérature.

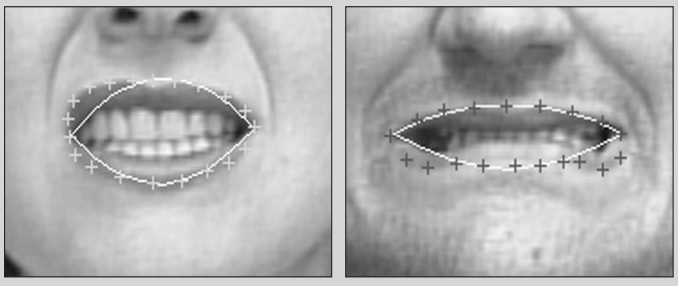


Fig. 2.14. Le modèle bi-parabolique est trop simple pour représenter les formes asymétriques. Les croix représentent les contours réels estimés par les auteurs, d'après [Tian *et al*., 2000].

rieur et 4 pour l'intérieur (voir figure 2.13-b). Les côtés gauche et droit de la bouche étant modélisés séparément, la représentation des contours asymétriques est améliorée. Cependant, comme Pantic suppose que les points d'annulation des dérivées des paraboles sont tous sur le même axe, *l'arc de Cupidon* ne peut être correctement représenté et le positionnement du contour supérieur reste approximatif. Dans [Coianiz *et al.*, 1995], Coianiz propose également de modéliser le contour supérieur par 2 paraboles (voir figure 2.13-c), mais les points d'annulation des dérivées sont laissés libres. Cela permet d'esquisser la forme de l'arc de Cupidon comme pour le modèle de Yuille, mais avec une complexité algorithmique réduite (car les paraboles sont contrôlées par moins de paramètres que les quartiques). Le modèle de Coianiz est d'ailleurs repris par Faruquie *et al.* [Faruquie et al, 2000], et est complété par 2 paraboles pour représenter le contour intérieur. Une autre solution pour le contour supérieur est de le modéliser par une seule quartique [Wark and Sridharan, 1998]. La quartique ayant une dérivée qui peut théoriquement s'annuler 3 fois, elle peut reproduire (au moins grossièrement) l'arc de Cupidon (voir figure 2.13-d). De plus, 5 paramètres suffisent à la contrôler, au lieu de 10 dans le cas de 2 quartiques et de 6 pour 2 paraboles. Cette solution offre donc un bon compromis entre le modèle très simple (mais approximatif) de la figure 2.13-a et ceux, plus élaborés, des figures 2.12 et 2.13-c. Cependant, le fait de remplacer la parabole supérieure par une quartique n'apporte bien souvent qu'une amélioration "esthétique" (dans le sens où le modèle ressemble un peu plus à une bouche) car le nombre de paramètres de contrôle n'est pas suffisant pour assurer à la fois le suivi précis de *l'arc de Cupidon* et celui des contours adjacents. Afin d'obtenir une grande déformabilité tout en conservant un nombre restreint de paramètres, Liew *et al.* utilisent le modèle de Coianiz (figure 2.13-c) en forçant les 2 paraboles supérieures à être symétriques [Liew *et al.*, 2000]. Puis, de manière à augmenter le nombre de degrés de liberté, ils effectuent 2 transformations : une torsion globale du modèle et un aplatissement de la parabole inférieure (voir figure 2.13-e). Au final, ils obtiennent un modèle aussi simple que celui de Coianiz du point de vue calculatoire, mais offrant un rendu beaucoup plus fidèle du contour des lèvres.

2.4.1.3 La déformation du modèle

Le modèle proposé par Yuille est contrôlé par un nombre important de paramètres. De plus, le calcul de l'énergie qui lui est associée est relativement complexe et fait intervenir de nombreux coefficients obtenus de manière heuristique. Dès lors, la descente de gradient qui permet d'assurer la convergence est longue et a de fortes chances de mener à un minimum local d'énergie. La première tentative de simplification de la méthode de Yuille est attribuée à Hennecke [Hennecke *et al.*, 1994]. Il constate tout d'abord qu'une partie importante du temps de calcul est consacrée à la détermination de l'énergie externe et des champs qu'elle utilise. Il propose donc de supprimer les champs *vallée* et *pic*, et ne garde qu'une version simplifiée du champ *contour.* Comme les contours de la bouche sont principalement horizontaux, Hennecke n'utilise que la composante verticale du gradient. L'énergie externe contient donc un seul terme (au lieu de 3 pour Yuille) qui est une somme pondérée d'intégrales sur des scalaires (au lieu de

vecteurs pour Yuille) le long des 5 courbes du modèle. Ensuite, il simplifie également l'expression de l'énergie interne en supprimant l'énergie de symétrie (qui favorise l'aspect symétrique du modèle) et l'énergie de centrage. Pour cela, il utilise un modèle symétrique et centré par construction, ce qui permet à la fois de supprimer quelques paramètres et de rendre les énergies de symétrie et de centrage inutiles. L'expression de l'énergie interne qu'il propose est donc finalement très simple. Chacun des paramètres du modèle est associé à une *fonction coût* qui augmente rapidement lorsque sa valeur s'approche de bornes déterminées de manière heuristique. De plus, une contrainte de continuité temporelle de l'épaisseur des lèvres est également inclus dans l'énergie interne. L'optimisation des paramètres est ensuite effectuée par une *descente de gradient*. Hennecke affirme que la complexité algorithmique de sa méthode est "raisonnable", mais il ne donne aucun temps de calcul. Cependant, il est peu probable que la convergence puisse être effectuée en temps réel sur une séquence vidéo car, même après simplification, le nombre de paramètres à ajuster reste très important (le modèle de Hennecke en comporte 12 !). De plus, cette méthode nécessite de déterminer précisément de nombreux paramètres heuristiques. Rien que pour l'énergie interne, il faut établir les 12 domaines de variation des paramètres et les 2 coefficients de raideur pour la contrainte de continuité temporelle. De même, 7 coefficients de pondération doivent être fixés pour calculer l'énergie externe. Finalement, la méthode de Hennecke est intéressante car elle démontre qu'une version simplifiée de l'algorithme de Yuille permet encore d'effectuer un suivi relativement correct du contour des lèvres tout en diminuant (probablement) le temps de calcul. Mais le modèle utilisé reste trop complexe pour être déformé par une *descente de gradient*.

Une autre solution pour accélérer la convergence est d'utiliser une méthode d'optimisation itérative moins complexe. Dans [Liew *et al.*, 2000], Liew *et al.* utilisent l'information chromatique de l'image pour effectuer une classification floue non supervisée (voir section 2.2.3) du visage en deux zones : *lèvre* et *non-lèvre*. Ils obtiennent ainsi une carte de probabilité d'appartenance à la classe *lèvre*. Le modèle est alors positionné en trouvant l'ellipse qui sépare au mieux les classes *lèvre* et *non-lèvre*. De cette manière, l'orientation, la largeur et la position horizontale de la bouche sont déterminées très rapidement. Les 6 autres paramètres nécessaires sont ensuite obtenus par la technique du *gradient conjugué* (qui converge théoriquement plus rapidement que *la descente de gradient*). De plus, l'énergie du modèle est calculée très simplement puisqu'elle ne contient qu'un terme d'énergie externe. Cette dernière n'est pas liée aux pics, vallées ou contours de l'image, mais est obtenue par un calcul de probabilité. Plus le modèle sépare nettement les classes lèvre et non-lèvre, et plus l'énergie est faible. Au final, cette utilisation séquentielle de 2 méthodes d'optimisation associée à un calcul d'énergie très simple permet de faire converger le modèle beaucoup plus rapidement que Yuille ou Hennecke. Pour une image de taille 70×100, la segmentation est effectuée en moins de 0.1 seconde sur un PIII 500 MHz. Une méthode du même type est utilisée par Rao et Mersereau dans [Rao and Mersereau, 1995]. Une carte de probabilité est utilisée pour positionner et déformer un modèle à deux paraboles (voir figure 2.15).

Les algorithmes les plus rapides sont probablement ceux qui n'utilisent pas de techni-

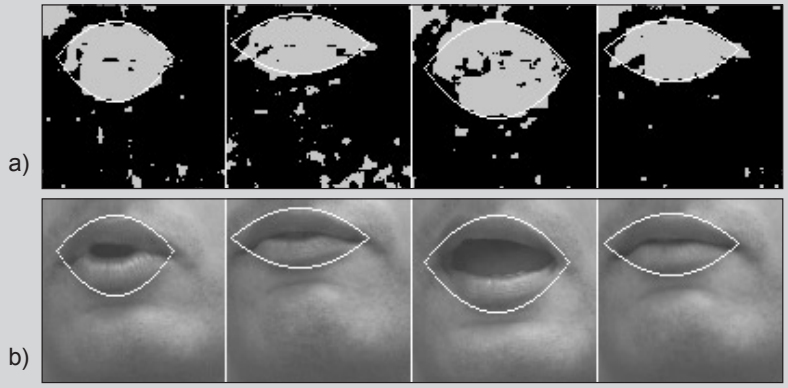
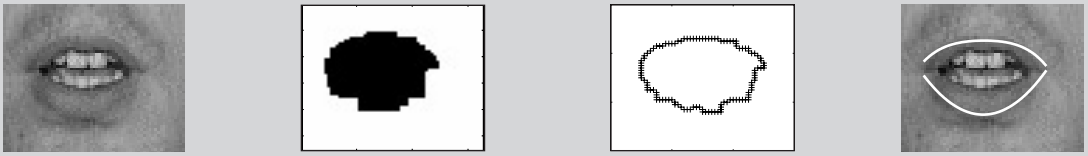


Fig. 2.15. a) Carte de probabilité d'appartenance aux lèvres, b) modèle bi-parabolique ajusté aux contours des lèvres, d'après [Rao and Mersereau, 1995].

ques d'optimisation itératives. Par exemple, dans [Wark and Sridharan, 1998], Wark effectue un seuillage du quotient R/G suivi d'opérations morphologiques pour obtenir un masque binaire de la bouche (voir section 2.2.2). Les points de contour extraits de ce masque sont ensuite utilisés pour calculer les 2 courbes du modèle par la méthode des *moindres carrés* (voir figure 2.16). Cette technique d'optimisation n'étant pas itérative, la déformation du modèle peut se faire très rapidement. Les *moindres carré*s sont également utilisés dans [Botino, 2002] pour calculer les paraboles supérieure et inférieure du modèle de la figure 2.13-a. Mais, dans ce cas seuls 8 points caractéristiques du contour sont considérés, ce qui accélère encore l'algorithme. La position de ces points est déduite de la position dans l'image précédente par l'algorithme de suivi de Lucas-Kanade. Dans [Tian *et al*., 2000], Tian *et al*. décrivent une méthode similaire, à la différence près que le nombre de points caractéristiques suivis d'une image à l'autre se réduit à 4 (les 2 commissures et les 2 extrema verticaux). Dès lors, comme une parabole est entièrement définie si 3 de ses points sont connus, les *moindres carrés* ne sont plus nécessaires et le calcul



a) image de départ | b) masque binaire des lèvres | c) points de contour | d) courbes du modèle

Fig. 2.16. Détermination du modèle optimal par les *moindres carrés*, d'après [Wark and Sridharan, 1998].

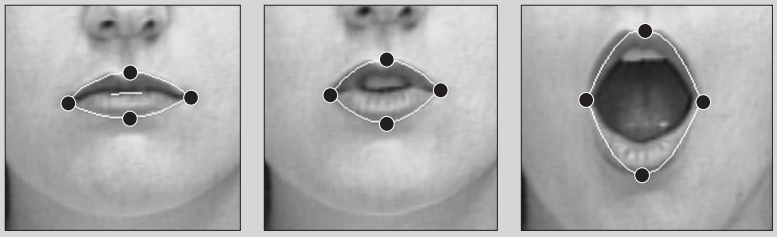


Fig. 2.17. Utilisation de 4 points caractéristiques (●) pour calculer les paraboles d'un modèle de lèvres, d'après [Tian *et al.*, 2000].

du modèle se résume à 2 inversions de systèmes matriciels de taille 3. Ces méthodes basées sur l'utilisation de points caractéristiques (dont quelques résultats sont présentés à la figure 2.17) sont probablement les plus rapides, mais elles présentent 2 inconvénients. En premier lieu, elles n'utilisent que la position de quelques points (i.e. des informations *locales*) pour calculer le modèle. Elles ne bénéficient donc pas de la même robustesse vis-à-vis du bruit que les méthodes utilisant des intégrales le long de courbes (i.e. des informations *globales*). Ensuite, elles reposent entièrement sur des algorithmes de suivi de point. Or, la précision de ces derniers est fortement diminuée lorsque les points sont situés dans des zones très déformables ou sur des contours mal définis. Dans le cas des lèvres, nos expérimentations ont montré que, sous éclairage naturel et sans maquillage, les erreurs de suivi devenaient très importantes après quelques images. Mais nous nous étendrons plus longuement sur ce dernier point dans le chapitre 4.

2.4.2 Les formes actives

Les *formes actives* (ou *ASM - Active Shapes Method*) ont été initialement proposées par Cootes et Taylor [Cootes and Taylor, 1992][Cootes *et al.*, 1995] pour localiser des objets déformables dans des images médicales. Rapidement, elles ont été étendues à de nombreux domaines, dont l'analyse labiale. Elles permettent également d'intégrer des connaissances a priori sur les formes admissibles dans un processus de segmentation. Cependant, contrairement aux *modèles déformables*, ces connaissances ne sont pas heuristiques mais statistiques. Le modèle utilisé ainsi que les déformations qu'il subit sont issus de l'analyse d'un grand nombre d'images d'entraînement.

2.4.2.1 Algorithme de base

Les formes actives sont des modèles statistiques représentant les contours d'un objet à segmenter. Elles s'appuient sur un *modèle de distribution de points* (*Point Distribution Model - PDM*) obtenu par entraînement. Pour cela, une base suffisamment représentative est étiquetée manuellement. Sur chaque image, N points (généralement quelques dizaines) sont positionnés

sur les contours. Ainsi, le contour de l'image *i* est représenté par le vecteur suivant :

$$\boldsymbol{x_i} = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, ..., x_{iN}, y_{iN})^T \qquad \text{(eq. 2.19)}$$

Puis, un algorithme itératif est utilisé pour translater, tourner et mettre à l'échelle les ensembles de points x_i de manière à les aligner. Cet alignement permet de supprimer les variations dues à la pose et de ne garder que celles reflétant des changements de forme. Une *analyse en composantes principales* (*Principal Components Analysis - PCA*) sur l'ensemble d'entraînement aligné donne ensuite la position moyenne des points (c'est-à-dire la forme de bouche moyenne) ainsi que les axes de déformation représentant au mieux les écarts à la moyenne. La forme de bouche moyenne peut être exprimée de la manière suivante :

$$\bar{\boldsymbol{x}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{x}_i \qquad \text{(eq. 2.20)}$$

où *M* est le nombre de bouches de la base d'entrainement. De plus, toute forme peut être approchée par une somme pondérée des modes principaux de l'*analyse en composantes principales* :

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \mathbf{Pb} \qquad \text{(eq. 2.21)}$$

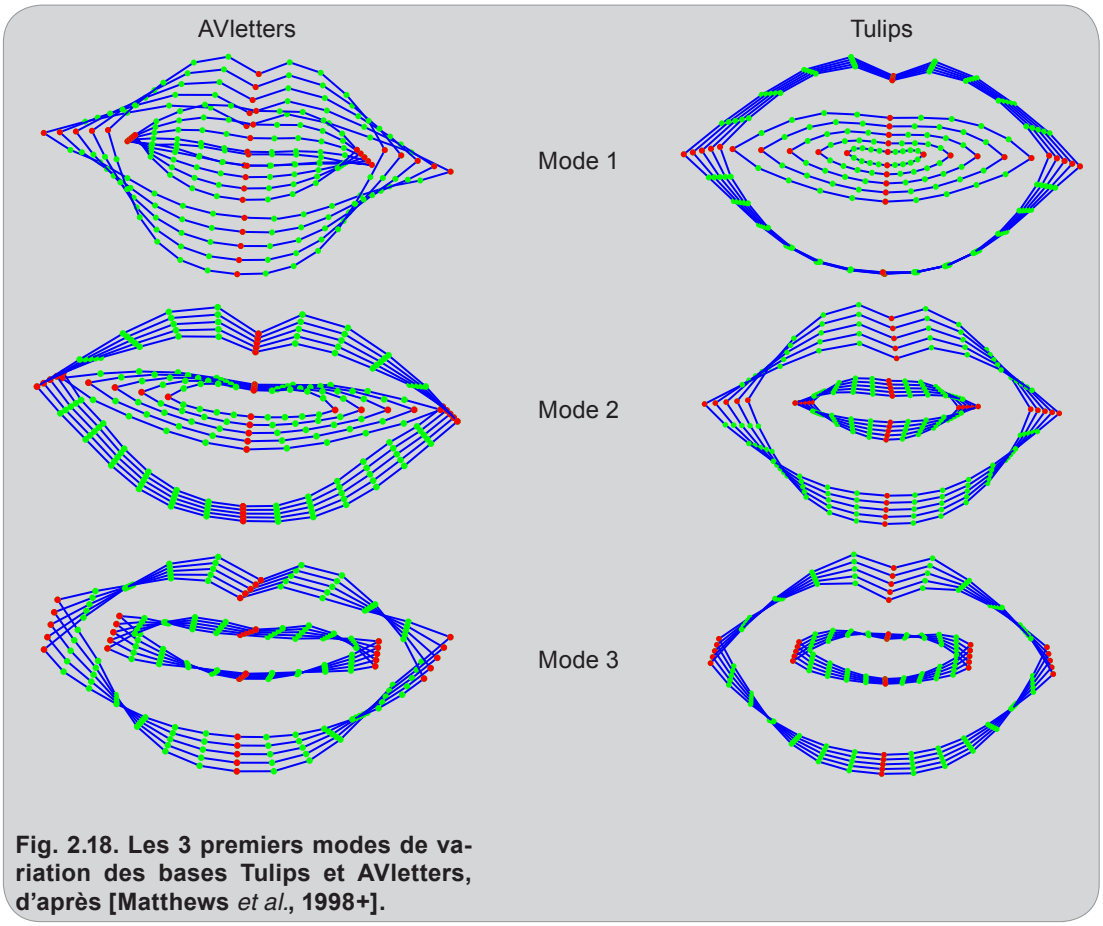


Fig. 2.18. Les 3 premiers modes de variation des bases Tulips et AVletters, d'après [Matthews *et al.*, 1998+].

où $\boldsymbol{P}=(p_1,p_2,...,p_t)$ est la matrice des t modes de variation les plus significatifs, et $\boldsymbol{b}=(b_1,b_2,...b_t)^T$ un vecteur de pondération. En général, les b_k peuvent varier entre $-3\sqrt{\lambda_k}$ et $3\sqrt{\lambda_k}$, où λ_k est la variance du mode k. Le nombre de vecteurs propres t est choisi de manière à ce que 95% de la variance observée lors de l'entraînement soit reproductible par l'équation 2.21. Comme t est en général bien inférieur au nombre de points du modèle, l'*ACP* permet d'obtenir une forte réduction du nombre de dimensions du problème. La figure 2.18 présente les 3 premiers modes de variation sur les bases *Tulips* et *AVletters*. Les 2 premiers modes de *AVletters* représentent les ouvertures verticale et horizontale respectivement. Pour *Tulips*, ces 2 modes sont inversés. Le troisième mode représente pour les 2 bases la déformation caractéristique du sourire.

Une fois que l'entraînement a été effectué, le contour des lèvres est localisé par un algorithme itératif :

1 - lors de la phase d'initialisation, le modèle moyen est positionné approximativement ;

2 - une recherche perpendiculaire au modèle permet de rapprocher les points des contours de l'image en les positionnant sur un maximum de gradient ;

3 - détermination du vecteur de pondération $\boldsymbol{b}$ qui permet d'approximer au mieux la distribution de points obtenue à l'étape 2. Pour cela, il suffit de résoudre :

$$\mathbf{b} = \mathbf{P}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \qquad \text{(eq. 2.22)}$$

4 - retourner à l'étape 2 jusqu'à la convergence.

Le succès de la méthode est dépendant de la qualité de la base d'images sur laquelle on effectue l'apprentissage. Elle doit être correctement normalisée afin d'assurer la convergence de l'algorithme [Belhumeur *et al.,* 1991]. De plus, l'analyse est très sensible aux conditions d'éclairage et aux variations de pose du visage. Enfin, comme la déformation est basée sur la recherche de maxima de gradient, la technique des *formes actives* conduit souvent à des résultats très approximatifs si les contours sont mal dessinés [Luettin *et al*., 1996].

2.4.2.2 Raffinements

Utilisation des niveaux de gris (GLDM)

De manière à rendre les *formes actives* plus robustes, quelques "raffinements" ont été proposés. En plus du modèle de forme, Cootes propose d'utiliser un *modèle de distribution des niveaux de gris* (*Grey Levels Distribution Model - GLDM*) [Cootes *et al*., 1993]. Cela permet de rechercher la forme optimale en tenant compte non seulement des informations de gradient, mais aussi des informations de luminance. Pour constituer le *GLDM*, les profils de niveaux de gris perpendiculaires au contour sont extraits et concaténés (voir figure 2.19). Ainsi, pour chaque image i, un vecteur global des niveaux de gris est construit :

$$\boldsymbol{h}_i = (gp_{i1}, gp_{i2}, ..., gp_{iN})^T \qquad \text{(eq. 2.23)}$$

où gp_{ij} est le profil correspondant au point j dans l'image i. Puis, une *analyse en composantes principales* permet de calculer la distribution moyenne $\bar{\boldsymbol{h}}$ ainsi que la matrice des modes principaux de variation $\boldsymbol{P}_{gp}$. Ainsi, tout profil peut être approximé par la combinaison suivante :

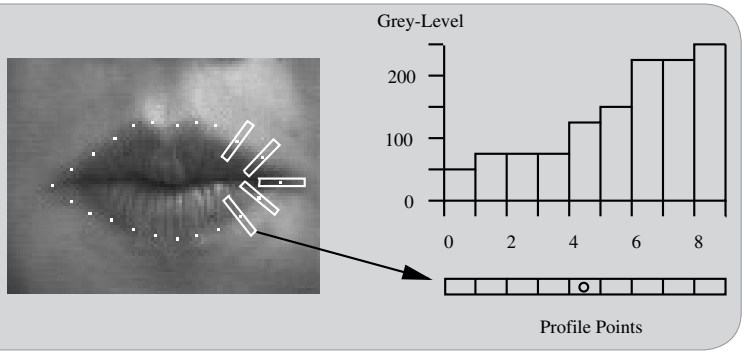


Fig. 2.19. Construction d'un modèle de distribution des niveaux de gris, d'après [Luettin, 1997].

$$\boldsymbol{h} = \bar{\boldsymbol{h}} + \mathbf{P}_{\mathbf{gp}}\mathbf{b}_{\mathbf{gp}} \qquad \text{(eq. 2.24)}$$

où $\boldsymbol{b}_{gp}$ est un vecteur de pondération. La figure 2.20 présente les 3 modes principaux de variation sur la forme moyenne. En toute rigueur, les modes sont représentables par un ensemble de profils unidimensionnels disposés autour du modèle de forme. Cependant, pour une question de lisibilité, il est courant d'effectuer une interpolation entre les profils pour obtenir une image continue (comme sur la figure 2.20). Le premier mode rend compte des changements d'illumination globale, le second décrit principalement l'éclairement de la lèvre inférieure, et le troisième reflète le contraste entre la peau et les lèvres [Luettin, 1997]. Les modes suivants permettent de caractériser des variations plus fines comme la direction de l'éclairage, la spécularité, la présence de la langue et des dents...

L'ajustement du modèle au contour de l'image se fait ensuite par un algorithme assez proche de celui des *formes actives* classiques :

1 - positionnement approximatif du modèle lors de l'initialisation ;

2 - calcul des profils de niveaux de gris orthogonaux au contour et du vecteur $\boldsymbol{b}_{gp}$ permettant de les approximer au mieux ;

3 - calcul de la distance D entre le profil observé *h* et son approximation :

$$D = (\boldsymbol{h} - \bar{\boldsymbol{h}})^T(\boldsymbol{h} - \bar{\boldsymbol{h}}) - \mathbf{b}_{\mathbf{gp}}^T\mathbf{b}_{\mathbf{gp}} \qquad \text{(eq. 2.25)}$$

4 - déformation du *PDM* (souvent en utilisant la méthode du *simplex*) de manière à diminuer *D ;*

5 - retour à l'étape 2 jusqu'à la convergence.

Cette méthode a l'avantage de pouvoir être applicable même si les contours sont mal dessinés car elle est basée sur la minimisation d'une fonction coût *D* qui n'utilise pas l'information de gradient.

Modélisation de l'apparence (AAM)

La modélisation de l'apparence, très «en vogue» depuis quelques années, repose sur les mêmes bases théoriques que les ASM. Introduite également par Cootes [Cootes *et al.*, 1998], cette technique connue sous le nom de *modèles d'apparence active* (*Active Appearence Models - AAM*) est en fait une généralisation des *formes actives*. Elle permet de construire un modèle

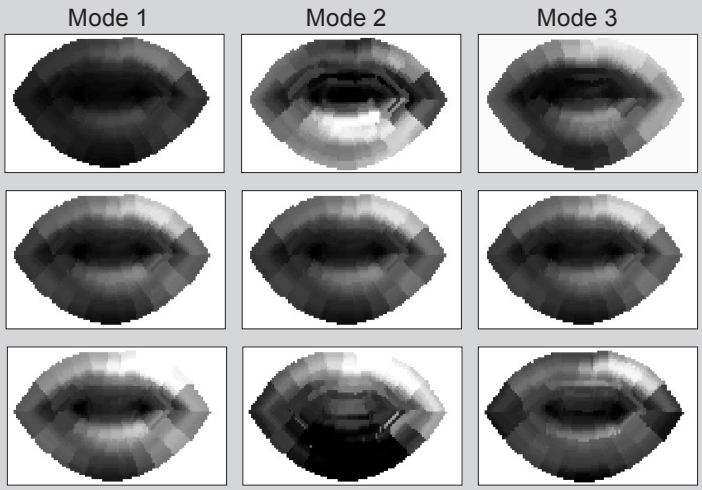


Fig. 2.20. Les 3 modes principaux de variation des niveaux de gris obtenus sur la base *Tulips*, d'après [Luettin, 1997].

statistique de l'objet à segmenter incluant à la fois la forme et les niveaux de gris. Cependant, contrairement à la méthode décrite précédemment, les niveaux de gris sont mesurés sur toute la forme, et non pas sur quelques profils perpendiculaires au modèle.

La technique des *AAM* nécessite le même type d'apprentissage que les *ASM*. Un certain nombre de points caractéristiques sont placés manuellement sur des images d'entraînement. La forme moyenne ainsi que ses principaux modes de variation sont ensuite calculés par *ACP*. Pour constituer le *modèle de luminance*, les images d'entraînement sont préalablement déformées par un algorithme de triangulation de manière à mettre en coïncidence les points du modèle de forme avec ceux de la forme moyenne. Puis, les niveaux de gris des pixels situés à l'intérieur du contour sont échantillonnés. Une analyse en composantes principales permet ensuite de calculer les modes principaux de variation de la luminance. Enfin, le *modèle d'apparence* est obtenu en concaténant les paramètres de forme et de luminance de chaque image. La figure 2.21 présente les 3 premiers modes obtenus par un entraînement sur la base *AVletters*. On peut observer qu'ils reflètent non seulement des changements de forme, mais également de luminance.

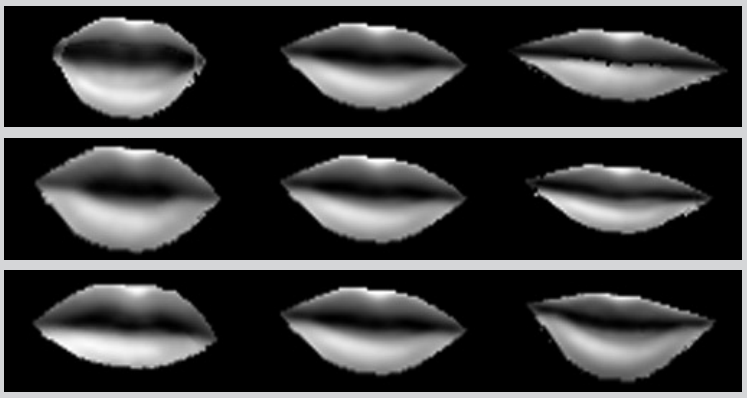


Fig. 2.21. Les 3 premiers modes du modèle d'apparence des lèvres obtenus sur la base *AVletters*, d'après [Matthews *et al.*, 1998]

La convergence du modèle d'apparence se fait par un algorithme itératif. Les paramètres sont progressivement ajustés de manière à minimiser la distance entre l'image de départ et l'image synthétisée par combinaison linéaire. Sur la figure 2.22, on peut remarquer que, lors de la convergence, la forme et la luminance sont modifiées en même temps.

La technique des *modèles d'apparence* présente quelques avantages incontestables. Tout d'abord, elle ne nécessite pas le réglage fin de paramètres par un expert (comme c'est le cas pour les *contours actifs*) ni la construction d'un modèle analytique spécifique à chaque forme (comme pour les *modèles déformables*). Ensuite, elle est applicable à n'importe quelle classe d'objet (visage, cœur, voiture...) puisqu'elle repose sur un apprentissage statistique effectué sur des exemples d'objet de la classe. Enfin, elle permet d'effectuer des approximations par synthèse quasiment photo-réalistes (voir figure 2.22).

L'apprentissage statistique est la grande force des AAM, mais c'est aussi l'origine de nombreux inconvénients. Pour une simple question matérielle, il est souvent difficile de constituer les modèle d'entraînement. Par exemple, dans [Matthews *et al.*, 1998+], Matthews a dû placer plus de 9800 points pour étiqueter 223 images de la base *Tulips* et plus de 20000 points pour la base *AVletters* ! Bien sûr, il est possible de réduire le nombre d'images d'entraînement ou le nombre de points du modèle ; mais, dans ce cas, l'analyse statistique devient peu représentative et les modes de déformations extraits ne permettent de couvrir qu'une faible variabilité de formes. Ensuite, le nombre de modes considérés doit permettre de représenter au moins 95% de la variance observée lors de l'apprentissage. Or, le fait d'inclure la luminance dans le modèle statistique augmente cette variance de manière très significative et oblige à utiliser beaucoup plus de modes que pour les *ASM*. Dans [Matthews *et al.*, 1998+], les formes de bouche des bases *Tulips* et *AVletters* sont décrites par 6 modes principaux. Dans [Matthews *et al.*, 1998], les *AAM* sont utilisées sur les mêmes bases et nécessitent 37 modes d'apparence. Cela ralentit inévitablement le processus de convergence [Cootes *et al.*, 1999]. Enfin, l'inconvénient le plus gênant des *apparences actives* est probablement le fait qu'elles synthétisent des apparences nouvelles par combinaison linéaire d'apparences (ou de modes) connues. Ainsi, tout changement structurel (i.e. opération hautement non-linéaire) non représenté dans la base d'entraînement ne pourra être reproduit. Par exemple, les AAM ne pourront pas créer un visage dont l'un des yeux

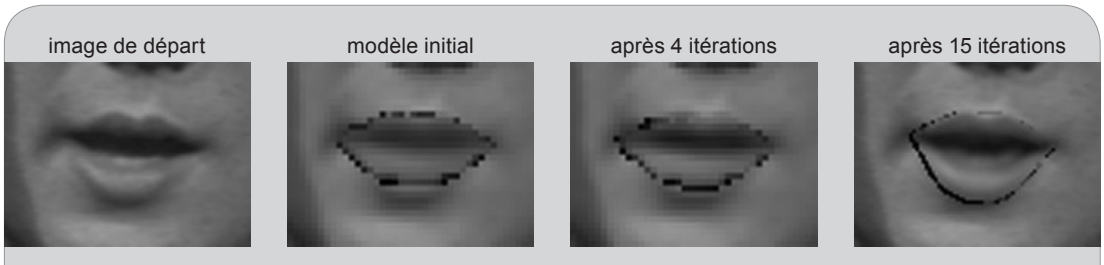


Fig. 2.22. Convergence du modèle d'apparence, d'après [Matthews *et al.*, 1998]

est fermé si aucune image de la base ne contient une telle configuration. Si le système doit pouvoir gérer les clins d'œil, il faut obligatoirement que l'entraînement inclut un nombre suffisant de clins d'œil. De même, l'apparition des dents ne pourra être reproduite si la base de départ ne contient que des bouches fermées. Bref, si on veut que le système puisse suivre les mouvements faciaux et labiaux, il faut utiliser une base d'apprentissage contenant toutes les configurations possibles. Cela rend le travail d'étiquetage fastidieux, d'autant plus que chaque configuration doit être suffisamment représentée sinon elle risque d'être considérée comme marginale. De plus, chaque nouvelle structure nécessitant l'ajout de modes propres supplémentaires, le nombre de paramètres à ajuster rend la convergence difficile.

2.5 Conclusion

Comme nous venons de le voir dans ce chapitre, les techniques de segmentation peuvent être regroupées en 3 grandes familles. Selon qu'elles utilisent des informations locales ou globales, on parle d'analyses de *bas niveau*, de *niveau moyen* et de *haut niveau*.

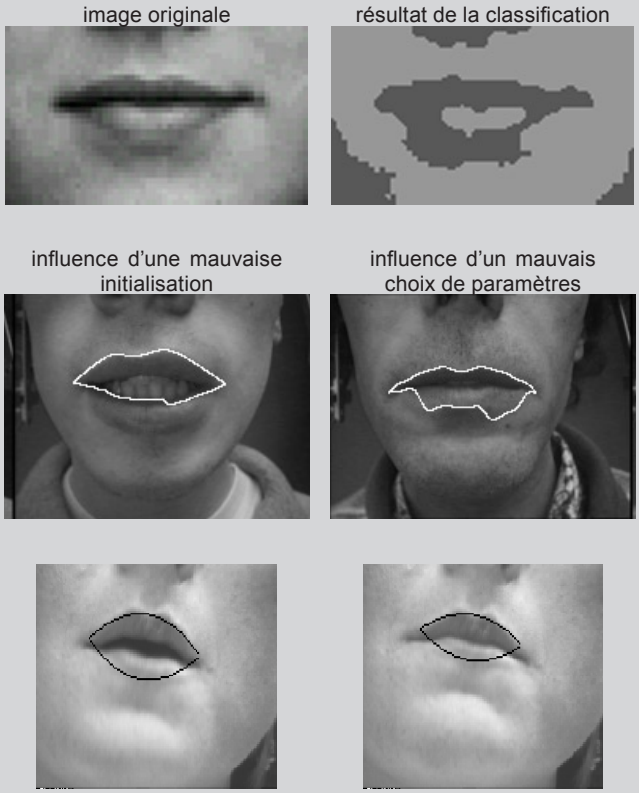
Les méthodes de bas niveau n'utilisent que les informations des pixels de l'image. Elles supposent que les pixels des lèvres possèdent des caractéristiques homogènes et différentes de la peau. Théoriquement, la segmentation peut donc être effectuée par l'identification et la séparation des classes *lèvre* et *peau*. En pratique, quel que soit l'espace dans lequel on se place, les distributions associées aux classes se chevauchent très souvent, ce qui conduit à des masques binaires très bruités. Certains algorithmes qui utilisent les *champs de Markov aléatoires* ou des traitements morphologiques permettent d'obtenir des masques plus compacts. Mais, comme aucune contrainte de régularité du contour n'est fixée, les frontières des zones segmentées sont souvent imprécises et bruitées (voir figure 2.23-a). Finalement, les méthodes de bas niveau permettent d'effectuer des localisations rapides et assez précises des zones d'intérêt, mais ne permettent absolument pas d'effectuer une détection fiable du contour des lèvres.

Les *snakes* utilisent également les informations contenues dans l'image (par l'intermédiaire de l'énergie extérieure), mais ils sont en plus soumis à des contraintes de régularité (par l'énergie intérieure). Des paramètres de courbure et d'élasticité leur donnent un aspect régulier et lisse. De plus leur grande déformabilité leur permet de s'adapter facilement à des formes très variées. Cette propriété est très intéressante lorsqu'il s'agit de segmenter des objets dont on ne peut prévoir à l'avance la forme (réseau de routes, vaisseaux sanguins, nuages,...). Mais elle apparaît plus comme un handicap si on s'intéresse à des objets dont la structure est connue (bouche, visage, main,...). Dans ce cas, la seule façon d'influencer le comportement des *contours actifs* est d'ajouter des contraintes heuristiques. Il est également possible de jouer sur les paramètres de courbure et d'élasticité, mais ce réglage est long (car basé sur des essais successifs) et est en général difficilement transposable à d'autres images. La figure 2.23-b montre le contour obtenu avec un jeu de paramètres mal adapté. Enfin, les *snakes* permettent difficilement de segmenter les zones de faible gradient comme les commissures des lèvres.

Contrairement aux *contours actifs*, les *méthodes de haut niveau* sont basées sur des modèles caractéristiques des lèvres, obtenus de manière heuristique ou statistique. Ainsi, la segmentation aboutit toujours à une forme admissible. De plus, comme la structure globale du modèle est conservée tout au long du processus de convergence, l'influence du bruit ou des perturbations locales est nettement moins importante que pour les *snakes* ou pour les *méthodes de bas niveau*. Cependant, dans le cas où le modèle est défini de manière heuristique (modèle déformables analytiques), il est difficile de trouver un compromis entre flexibilité, précision et complexité calculatoire. Un modèle trop simple convergera très vite, mais aura peu de degrés de liberté et ne pourra qu'approcher les contours des lèvres (voir figure 2.23-c). Pour résoudre ce problème, il est possible de construire le modèle et de déterminer ses modes de déformation par apprentissage statistique (*ASM* et *AAM*). Ainsi, la forme générique peut être relativement complexe et réaliste. Mais l'apprentissage est fastidieux et long si on veut couvrir une grande plage de variabilité. De plus, dans le cas d'objets pouvant comporter des occultations ou des apparitions, la base d'entraînement doit inclure des exemples (en nombre suffisant) de toutes les configurations possibles. Cela rend l'étiquetage plus long et augmente le nombre de modes principaux de variation. Enfin, les méthodes basées sur des modèles statistiques sont très sensibles aux conditions d'éclairage et à l'angle de prise de vue.

Dans les chapitres suivants, nous proposons une **méthode hybride** utilisant des "ingrédients" propres à chacune des 3 familles décrites précédemment. Cette méthode doit permettre d'extraire le contour labial externe à une **cadence proche du temps réel** et doit être suffisamment **robuste** vis-à-vis des conditions d'illumination et de cadrage. De plus, de manière à être utilisable dans des applications biométriques ou de reconnaissance d'émotions, le contour doit être extrait avec un maximum de **précision**.

Tout d'abord, nous adopterons le même postulat que les *techniques de bas niveau* en supposant que les lèvres peuvent être caractérisées (au moins grossièrement) par leur couleur. Nous définirons pour cela quelques grandeurs colorimétriques permettant de séparer au mieux les lèvres de la peau. Ensuite, nous utiliserons les *contours actifs* pour leur grande flexibilité. Cependant, nous leur apporterons quelques modifications pour augmenter leur zone de convergence et pour rendre plus intuitif le réglage de leurs paramètres. Ce nouvel outil permettra de localiser facilement la bouche dans la première image de la séquence. La robustesse de notre méthode sera assurée par l'utilisation d'un *modèle déformable analytique* de bouche ne nécessitant, par conséquent, aucun apprentissage. Afin de satisfaire aux contraintes de précision que nous nous sommes fixées, nous proposerons un modèle original suffisamment déformable pour pouvoir reproduire des formes très différentes, et suffisamment simple pour permettre une convergence rapide. Enfin, nous verrons dans le chapitre 4 que l'utilisation d'informations temporelles permettra d'accroître la robustesse et la rapidité de notre méthode.



a) les *méthodes de bas niveau* conduisent généralement à des contours approximatifs, d'après [Gordan *et al.*, 2002].

b) si l'initialisation est mauvaise ou si les paramètres sont mal réglés, le *snake* suit des mauvais contours, d'après [Delmas, 2000].

c) un modèle déformable analytique trop simple permet seulement d'approximer les contour des lèvres, d'après [Chen and Rao, 1998].

Fig. 2.23. Quelques artefacts classiques des différentes méthodes de segmentation.

CHAPITRE 3

Segmentation statique

3.1 Introduction

La méthode de segmentation que nous proposons s'applique à des séquences vidéo et comporte 2 étapes principales : l'*initialisation* et le *suivi*. Dans ce chapitre, nous traitons de l'initialisation, le suivi étant décrit en détail dans le chapitre suivant. Lors de cette phase, le contour externe des lèvres est extrait de la première image d'une séquence. Par conséquent, on parle ici de «segmentation statique» car aucune information temporelle n'est utilisée.

En premier lieu, dans la partie 3.2 nous analysons les mélanges chromatiques associés aux lèvres et à la peau. Après une comparaison des espaces couleurs utilisés couramment en analyse faciale, nous proposons d'utiliser une grandeur colorimétrique permettant d'effectuer une bonne séparation des lèvres et de la peau. De plus, nous introduisons un *gradient hybride* qui combine à la fois les informations de luminance et de chrominance, et qui facilite la localisation de la frontière supérieure des lèvres.

Dans la partie 3.3, nous utilisons ce *gradient hybride* pour guider la convergence d'un nouveau type de contour actif : le «jumping snake». Sa convergence est une succession de phases de croissance et de saut, ce qui lui permet de franchir les zones bruitées plus facilement que les *contours actifs* classiques. De plus, nous montrons que ses paramètres sont faciles à régler. Le *jumping snake* permet de localiser le contour supérieur des lèvres avec une précision suffisante pour extraire quelques points caractéristiques.

Dans la partie 3.4, ces points caractéristiques sont utilisés comme points d'ancrage d'un modèle paramétrique. Contrairement à la plupart des modèles proposés dans la littérature, ce modèle original, composé de courbes cubiques, est suffisamment déformable pour représenter fidèlement les spécificités de lèvres très différentes.

Enfin, dans la partie 3.5, quelques résultats représentatifs sont commentés et nous discutons des limites de la méthode proposée.

3.2 Choix de grandeurs colorimétriques

Le choix d'indices colorimétriques adaptés au problème est une étape cruciale de toute segmentation. D'une manière générale, les premiers algorithmes d'analyse d'images étaient bridés par les faibles performances des caméras et du matériel informatique. Ils utilisaient donc exclusivement la luminance... Aujourd'hui, la démocratisation des caméras couleur ainsi que l'augmentation des puissances de calcul permettent d'utiliser les différents plans colorimétriques de l'image. Dans cette partie, nous allons voir comment la prise en compte de la couleur permet d'identifier les lèvres plus facilement.

3.2.1 Analyse chromatique des lèvres et de la peau

De nombreuses études ont montré que l'utilisation de la couleur améliore significativement les performances des algorithmes d'analyse faciale. En effet, la peau est caractérisée plus par sa couleur que par sa luminance [Yang *et al.*, 1997]. Même pour des personnes différentes, ses caractéristiques chromatiques restent relativement constantes alors que l'illumination peut varier énormément. Ainsi, le point de départ de la plupart des méthodes d'analyse faciale est la détermination d'un espace couleur dans lequel la luminance et la chrominance sont exprimées séparément de manière explicite.

L'espace HSV est très utilisé pour caractériser la peau [Pantic *et al.*, 2001][Zhang and Mersereau, 2000][Zarit *et al.*, 1998] . Dans cet espace, la chrominance est décrite par les composantes H et S (teinte et saturation), et la luminance par la composante V. Dans [Zhang and Mersereau, 2000], les auteurs constatent que la teinte des lèvres est relativement constante et bien séparée de celle de la peau. Ils proposent donc de l'utiliser pour localiser les lèvres. Cependant, nos expérimentations ne nous ont pas permis d'arriver aux mêmes conclusions. En effet, comme il s'agit d'une grandeur angulaire définie entre 0 et 2π, les zones rouges sont en général très bruitées car elle sont situées aux bornes de l'intervalle de définition. Si l'objet à étudier est à dominante rouge (comme les lèvres, par exemple), il est possible de décaler la teinte de manière à déplacer la zone de discontinuité vers une couleur peu présente dans l'image. La figure 3.1-e présente le plan teinte décalé de l'image 3.1-a. On peut constater que, même après le décalage, la teinte reste bruitée. De plus, contrairement à ce que Zhang et Mersereau affirment, la teinte ne permet pas de faire ressortir clairement les lèvres. L'histogramme de la figure 3.1-f montre d'ailleurs que les distributions des teintes des lèvres et de la peau sont très mal séparées.

L'espace YCrCb est également très utilisé en analyse faciale, car il a été montré que la peau occupe une zone relativement restreinte dans le plan (Cr, Cb) [Tsapatsoulis *et al.*, 2000]. Mais, là encore, nos tests ont montré que la distribution associée aux lèvres est très souvent confondue avec celle de la peau. L'histogramme dans le plan (Cr, Cb) présenté à la figure 3.1-d montre qu'il est impossible de définir 2 zones distinctes correspondant d'une part aux pixels de peau et, d'autre part, aux pixels de lèvres. Dans [Hsu *et al.*, 2002], les auteurs introduisent une

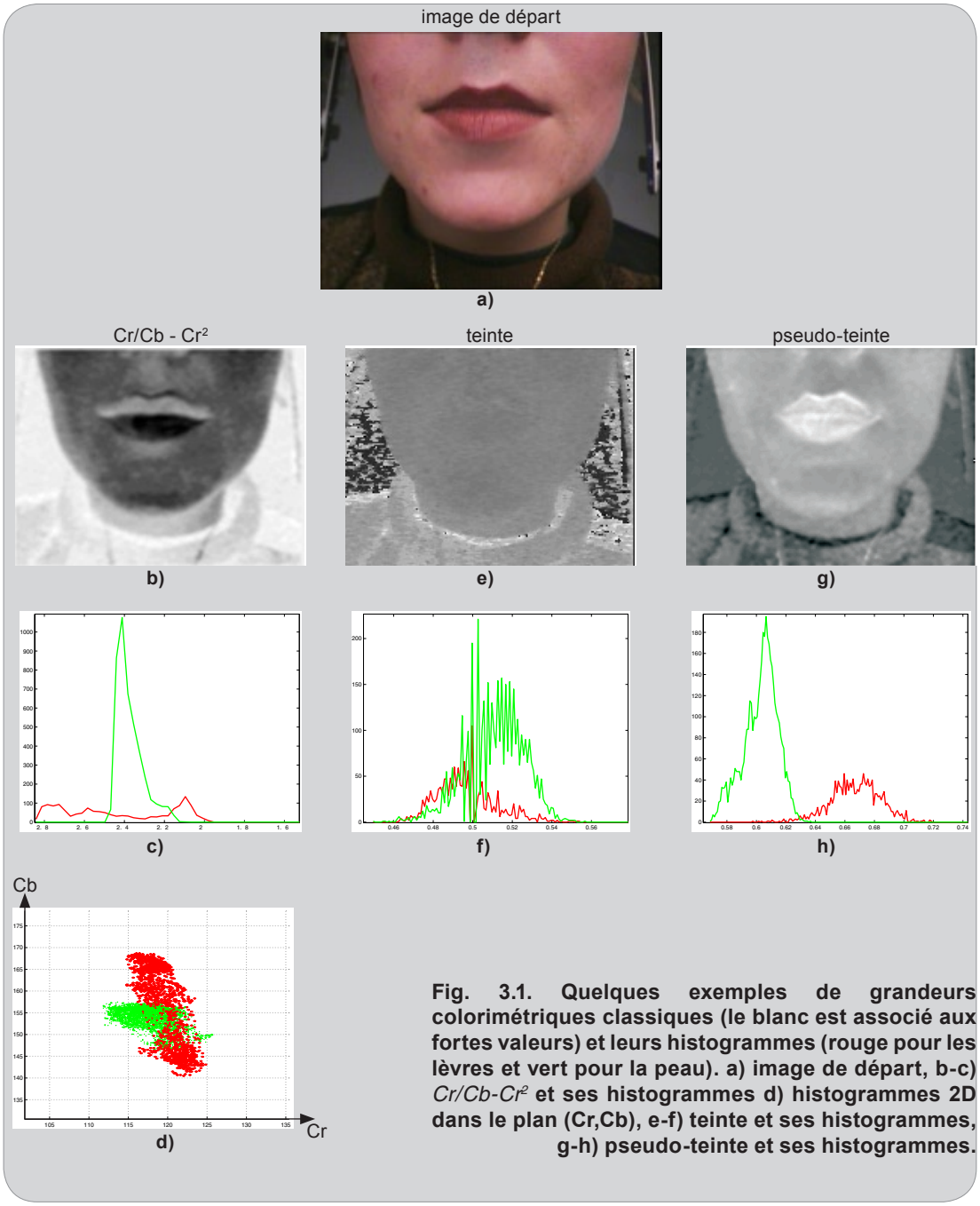


Fig. 3.1. Quelques exemples de grandeurs colorimétriques classiques (le blanc est associé aux fortes valeurs) et leurs histogrammes (rouge pour les lèvres et vert pour la peau). a) image de départ, b-c) *Cr/Cb-Cr²* et ses histogrammes d) histogrammes 2D dans le plan (Cr,Cb), e-f) teinte et ses histogrammes, g-h) pseudo-teinte et ses histogrammes.

nouvelle grandeur colorimétrique basée sur les composantes *Cr* et *Cb*. Ils calculent $Cr/Cb\text{-}Cr^2$ pour faire ressortir les lèvres et faciliter leur localisation. La figure 3.1-b présente le résultat de ce calcul appliqué à l'image 3.1-a. On constate que la valeur associée au lèvres n'est pas constante et qu'elle est parfois confondue avec celle de la peau (voir l'histogramme de la figure 3.1-c).

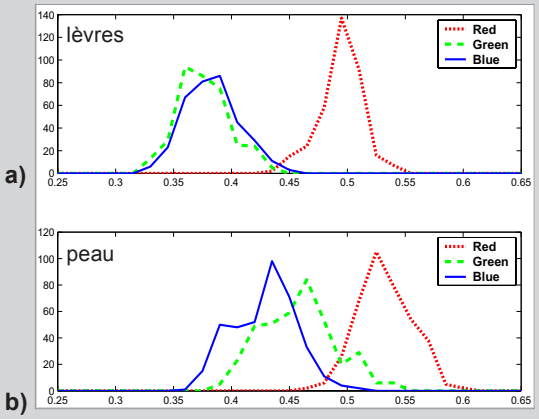


Fig. 3.2. Histogrammes typiques des composantes *R*,*G* et *B* pour les lèvres (a) et pour la peau (b).

Dans l'espace *RGB*, les lèvres et la peau ont des composantes différentes. Leurs histogrammes sont présentés à la figure 3.2. Dans les 2 cas, la composante rouge est prédominante. Cependant, il y a plus de vert que de bleu dans le mélange associé à la peau, alors que ces 2 composantes sont à peu près égales pour les lèvres. La peau apparaît généralement plus «jaune» que les lèvres car la différence entre ses composantes rouge et verte est moins importante que pour les lèvres. De nombreux auteurs ont utilisé cette propriété pour définir une grandeur colorimétrique caractéristique des lèvres. Par exemple, le quotient $Q=R/G$ a des valeurs plus importantes pour les lèvres que pour la peau. Il est utilisé pour localiser et segmenter les lèvres notamment dans [Chiou and Hwang, 1997][Wark and Sridharan, 1998] et [Lucey et al., 1999]. Cependant, comme lorsque G est faible le quotient R/G peut prendre des valeurs très importantes, Q est souvent bruité. Pour résoudre ce problème, Hulbert et Poggio ont proposé d'utiliser une pseudo-teinte bornée entre 0 et 1 [Hulbert and Poggio, 1998]:

$$h = \frac{R}{R+G} \qquad \text{(eq. 3.1)}$$

Contrairement à la teinte usuelle H, la pseudo-teinte est bijective. Elle est plus forte pour les lèvres que pour la peau (voir figure 3.1-g). De plus, l'histogramme de la figure 3.1-h montre que les distributions de pseudo-teinte des zones de lèvres et de peau sont bien séparées.

3.2.2 Gradient hybride

En plus de leur couleur spécifique, les lèvres possèdent une structure particulière et génèrent des zones d'ombre caractéristiques. Par exemple, si l'on suppose que le locuteur est éclairé par une source lumineuse placée au-dessus de lui, la frontière supérieure de la lèvre du dessus est une zone de forte luminance alors que la lèvre est elle-même dans une zone d'ombre (voir figure 3.3). Pour combiner ces informations de chrominance et de luminance, nous proposons d'utiliser le «gradient hybride» $\boldsymbol{R}_{top}$. Pour le pixel (x,y), il est calculé comme suit :

$$\boldsymbol{R}_{top}(x,y) = \nabla[h_N(x,y) - L_N(x,y)] \qquad \text{(eq. 3.2)}$$

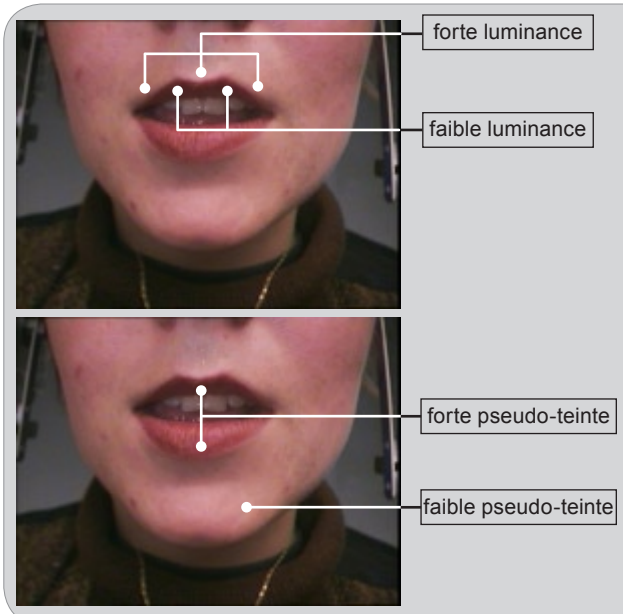


Fig. 3.3. Caractéristiques de luminance et de chrominance des différentes zones des lèvres.

où $\nabla[.]$ est l'opérateur gradient. $L_N(x,y)$ et $h_N(x,y)$ sont respectivement la luminance et la pseudo-teinte du pixel (x,y), normalisées entre 0 et 1 et à dynamique unitaire:

$$\begin{cases} h_N(x,y) = \dfrac{h(x,y) - \min(h)}{\max(h) - \min(h)} \\ L_N(x,y) = \dfrac{L(x,y) - \min(L)}{\max(L) - \min(L)} \end{cases} \qquad \text{(eq. 3.3)}$$

où *min(.)* et *max(.)* sont respectivement les minima et maxima calculés sur toute l'image. Il est à noter que cette normalisation est indispensable pour donner à la pseudo-teinte et à la luminance des valeurs comparables.

Comme le montre la figure 3.4, ce gradient hybride fait ressortir la frontière supérieure des lèvres bien mieux que les gradients de luminance ou de pseudo-teinte pris séparément. Dans la section suivante, il sera utilisé pour localiser la bouche.

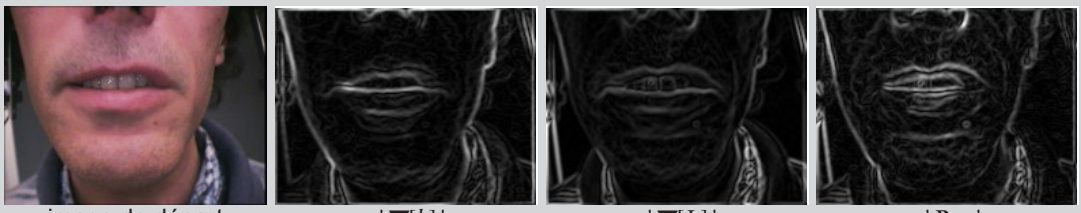


Fig. 3.4. Comparaison de différents types de gradients pour la localisation du contour supérieur de la bouche (le blanc est associé aux fortes normes).

3.3 Localisation de la bouche

L'algorithme de localisation que nous proposons ici repose sur la détection de quelques points caractéristiques des lèvres. Cette détection est effectuée en utilisant notamment un type de *contour actif* original possédant une zone d'initialisation relativement étendue : le *jumping snake*. Ce nouvel outil est présenté à la section suivante.

3.3.1 Le «jumping snake»

Les *contours actifs*, ou *snakes*, ont prouvé leur efficacité dans de nombreux problèmes de segmentation. Depuis leur apparition, de nombreuses améliorations ont été proposées dans la littérature. Mais aucune n'a vraiment réglé les problèmes de l'initialisation (qui doit être faite suffisamment près du contour final) et du réglage des paramètres (basé sur des essais successifs et rarement réutilisable sur d'autres images). La méthode présentée ici permet, dans une certaine mesure, de palier à ces inconvénients.

Pour localiser le contour supérieur de la bouche, nous utilisons un nouveau type de contour actif que nous appelons «*jumping snake*» (ou «serpent bondissant», pour les francophiles) car sa convergence est une succession de phases de croissance et de saut. Il est initialisé par un germe S^0 qui peut être situé relativement loin du contour final (l'étendue de la zone de convergence sera discutée à la section 3.5). Le germe est placé manuellement au-dessus de la bouche. Le jumping snake grandit ensuite à partir de ce point jusqu'à ce qu'il atteigne un nombre pré-déterminé de points. Cette phase de croissance est assez comparable au *growing snake* proposé par Berger et Mohr [Berger and Mohr, 1990] dans le sens où le contour est initialisé avec un seul point et est progressivement prolongé à chacune de ses extrémités. A la fin de la phase de croissance, le germe *saute* vers une autre position plus proche du contour final. Ce comportement discontinu permet de franchir plus facilement les zones bruitées. De plus, cela permet au *jumping snake* d'avoir une zone de convergence plus étendue. Le processus s'arrête lorsque l'amplitude du saut devient inférieure à un certain seuil. Le déroulement de l'algorithme est résumé sur la figure 3.5.

Durant la phase de croissance des points terminaux sont ajoutés aux extrémités droite et gauche. Ils sont situés à une distance horizontale constante (notée Δ) du point précédent (voir figure 3.6). De plus, la zone de recherche est restreinte au secteur angulaire $[\theta_{inf}, \theta_{sup}]$. Les meilleurs points terminaux, notés M_{i+1} et M_{-i-1}, sont trouvés dans cette zone en maximisant les flux moyens de $\boldsymbol{R}_{top}$ à travers les segments M_iM_{i+1} et $M_{-i-1}M_{-i}$. Ces flux sont calculés comme suit :

$$\begin{cases} \phi_{i+1} = \dfrac{\displaystyle\int_{M_i}^{M_{i+1}} \boldsymbol{R}_{top}.\boldsymbol{dn}}{|\,\boldsymbol{M_iM_{i+1}}\,|} \\ \phi_{-i-1} = \dfrac{\displaystyle\int_{M_{-i-1}}^{M_{-i}} \boldsymbol{R}_{top}.\boldsymbol{dn}}{|\,\boldsymbol{M_{-i-1}M_{-i}}\,|} \end{cases} \qquad \text{(eq. 3.4)}$$

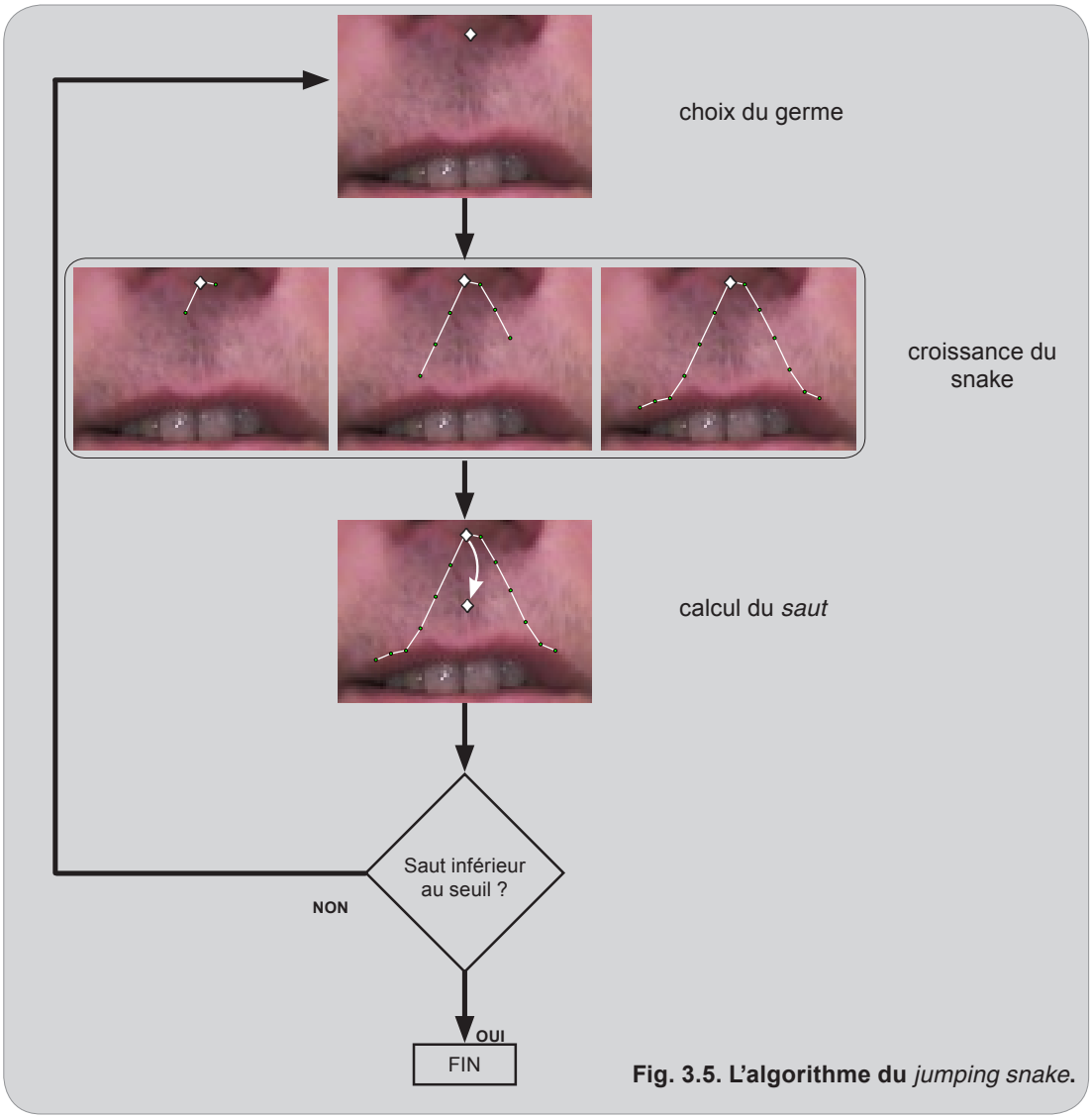


Fig. 3.5. L'algorithme du *jumping snake*.

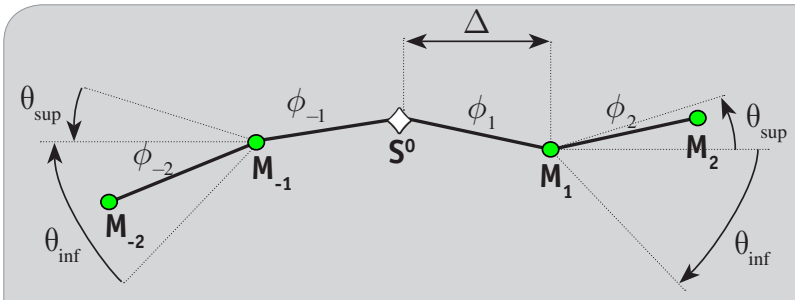


Fig. 3.6. A partir du germe S^0 (◇), le snake est allongé en ajoutant des points terminaux (●). Les flux moyens de R_{top} (ϕ_i) à travers chaque segment doivent être maximisés.

où ***dn*** est le vecteur orthogonal au segment considéré. Les maximisations des flux ϕ_{i+1} ϕ_{-i-1} sont effectuées en testant un nombre restreint de points candidats situés dans la zone de recherche. Ces points sont régulièrement espacés sur le segment inclus dans le secteur angulaire [θ_{inf}, θ_{sup}].

Lorsque le snake atteint un nombre pré-déterminé de points $2N+1$, la croissance s'arrête et la position du nouveau germe S^1 est calculée. Il s'agit de la **phase de saut** de l'algorithme. On note $\{M_{-N},...,M_{-1},S^0,M_1,...,M_N\}$ les points du snake à la fin de la croissance et $\{\phi_{-N},...,\phi_{-1},\phi_1,...,\phi_N\}$ les flux moyens à travers ses $2N$ segments. Le nouveau germe S^1 doit se rapprocher du contour final, c'est-à-dire des zones de fort gradient hybride $\boldsymbol{R}_{top}$. Il faut donc que S^1 se rapproche des segments dont le flux moyen est élevé. Pour cela, on considère que S^1 est le barycentre de S^0 et des points situés dans les zones de plus fort gradient. Sur la figure 3.7, ces points sont représentés par des gros cerles. Ils sont proches du contour final. Si l'on note $\{i_1,...,i_N\}$ les indices des points associés aux N plus forts flux moyens, alors la position verticale de S^1 s'écrit :

$$y_{S^1} = \frac{1}{2}\left(y_{S^0} + \frac{\sum_{k=1}^{N} \phi_{i_k}\, y(i_k)}{\sum_{k=1}^{N} \phi_{i_k}} \right) \qquad \text{(eq. 3.5)}$$

où $y(i_k)$ est la position verticale du point M_{i_k}. La position horizontale x_{s1} du germe est gardée constante.

Un nouveau snake grandit ensuite à partir de ce nouveau germe jusqu'à ce qu'il atteigne la longueur pré-déterminée et puisse *sauter* à nouveau. Le processus se répète tant que l'amplitude du saut est supérieure à 1 pixel. En général, la convergence est réalisée en 4 ou 5 sauts. Au final, les points du snake sont situés sur le contour supérieur des lèvres. Cependant, comme c'est un calcul de barycentre qui permet de déterminer la position du germe (et non pas la maximisation d'un flux), ce dernier a peu de chances de se trouver sur le contour. La figure 3.8 présente les position successives du snake au cours de la convergence. Les 5 germes sont symbolisés par des losanges (◇). On peut observer que le dernier germe (i.e. le plus bas des 5) n'est effectivement pas sur le contour supérieur des lèvres. Même si l'on répète le processus de *croissance-saut* un grand nombre de fois, le germe se rapproche très peu du contour et finit par se stabiliser à une position d'équilibre située généralement au-dessus des lèvres. Ainsi, si l'on

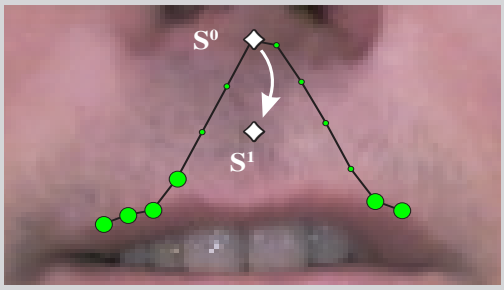


Fig. 3.7. La position du germe S^1 est calculée en utilisant S^0 et les points associés aux plus forts flux moyens (gros points).

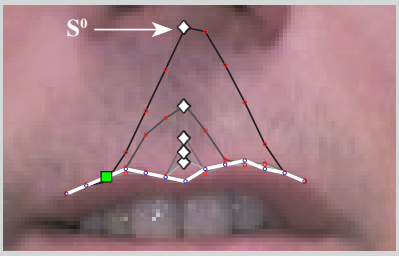


Fig. 3.8. Convergence du *jumping snake*. Après chaque saut, la position du nouveau germe (◇) est calculée. La dernière phase de croissance est effectuée à partir d'un germe situé sur le contour (■). Le snake final suit le contour supérieur des lèvres (ligne blanche).

souhaite que tous les points du snake soient correctement positionnés, il est nécessaire d'effectuer une dernière phase de croissance en utilisant un germe situé sur le contour. Pour cela, on choisit sur le dernier snake le point associé au flux moyen le plus fort. Sur la figure 3.8, ce point est représenté par un carré (■). La croissance effectuée à partir de ce germe est identique à celle décrite précédemment, à la différence près qu'elle s'arrête lorsque les bornes droite et gauches définies par les snakes précédents sont atteintes. Par conséquent, les nombres de points de chaque coté du germe peuvent être différents. La ligne blanche de la figure 3.8 représente le contour finalement obtenu.

Dans la partie suivante, nous allons voir comment le *jumping snake* est utilisé pour localiser quelques points caractéristiques des lèvres. Dans la partie 3.3.3, nous montrerons que cette méthode de détection impose quelques contraintes à partir desquelles il est possible de déterminer facilement la valeur des paramètres ($\theta_{inf}, \theta_{sup}, N, \Delta$) du *jumping snake*.

3.3.2 Détection des points caractéristiques hauts et bas

Les points caractéristiques donnent des indices importants sur la forme des lèvres. Ils sont utilisés comme points d'ancrage pour la construction du modèle paramétrique. De nombreuses méthodes basées sur l'utilisation de points caractéristiques ont déjà été proposées. Dans [Botino, 2002], Botino utilise 8 points caractéristiques régulièrement espacés sur le contour des lèvres pour modéliser les contours haut et bas par des paraboles (voir figure 3.9-a). Dans [Tian *et al.*, 2000], ces paraboles sont calculées en utilisant seulement 4 points caractéristiques (voir figure 3.9-b). La méthode que nous proposons nécessite 6 points caractéristiques principaux et 2 points secondaires (voir figure 3.9-c) :

- les 2 commissures (P_1 et P_5),
- les 3 points de l'arc de Cupidon (P_2, P_3 et P_4),
- le point bas (P_6),
- 2 points secondaires situés à l'intérieur de la bouche (P_7 et P_8).

Les 3 points de l'arc de Cupidon sont situés sur le contour détecté par le *jumping snake*. P_2 et P_4 sont les points les plus élevés de ce contour, de chaque coté de la verticale passant par le germe initial. P_3 est le point le plus bas du contour, entre P_2 et P_4 (voir figure 3.10).

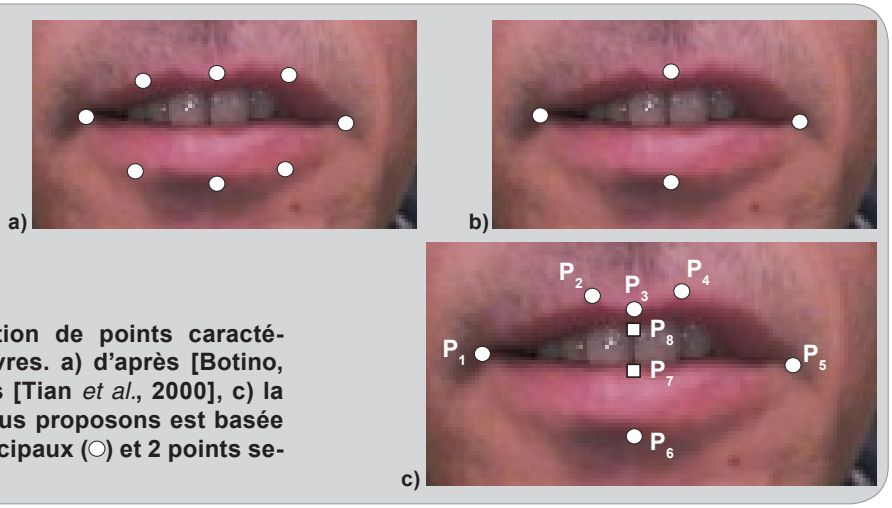


Fig. 3.9. Utilisation de points caractéristiques des lèvres. a) d'après [Botino, 2002], b) d'après [Tian *et al.*, 2000], c) la méthode que nous proposons est basée sur 6 points principaux (○) et 2 points secondaires (□).

Les point P_6, P_7 et P_8 sont trouvés en examinant $\nabla_y[h]$, la composante verticale du gradient de la pseudo-teinte, le long de l'axe vertical passant par le point P_3 (voir figure 3.10). La pseudo-teinte est plus importante pour les lèvres que pour la peau ou les dents. Ainsi, le point P_7 correspond au maximum de $\nabla_y[h]$ sous P_3. P_6 et P_8 sont associés respectivement au minimum de $\nabla_y[h]$ en-dessous et au-dessus de P_7. Il faut noter que, même lorsque la bouche est fermée, il est possible de repérer P_7 et P_8. En effet, dans la plupart des cas, la ligne sombre qui sépare les 2 lèvres fermées a une pseudo-teinte inférieure à celle des lèvres elles-même. Les extrema correspondant à P_7 et P_8 sont donc très rapprochés, mais parfaitement détectables, comme le montre la figure 3.10-b.

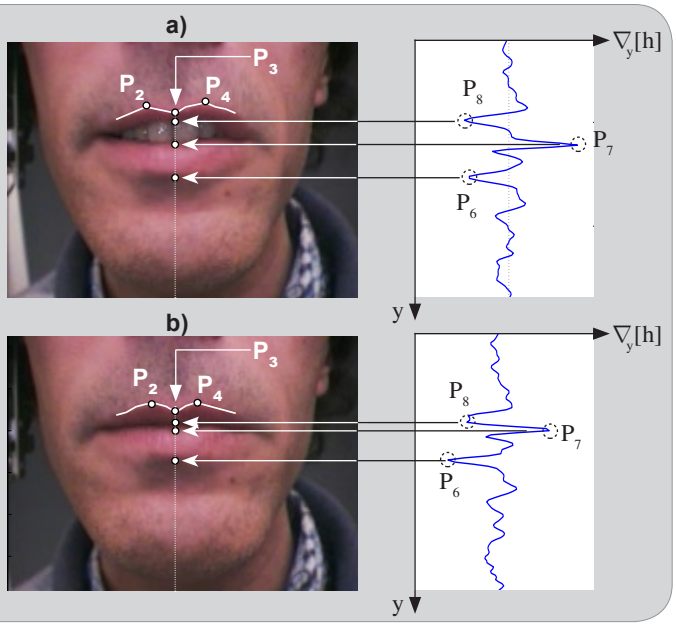


Fig. 3.10. Les 3 points supérieurs sont trouvés sur le contour détecté par le *jumping snake* (ligne blanche). P_6, P_7 et P_8 sont situés sous P_3, sur des extrema de $\nabla_y[h]$. a) Bouche ouverte, b) bouche fermée.

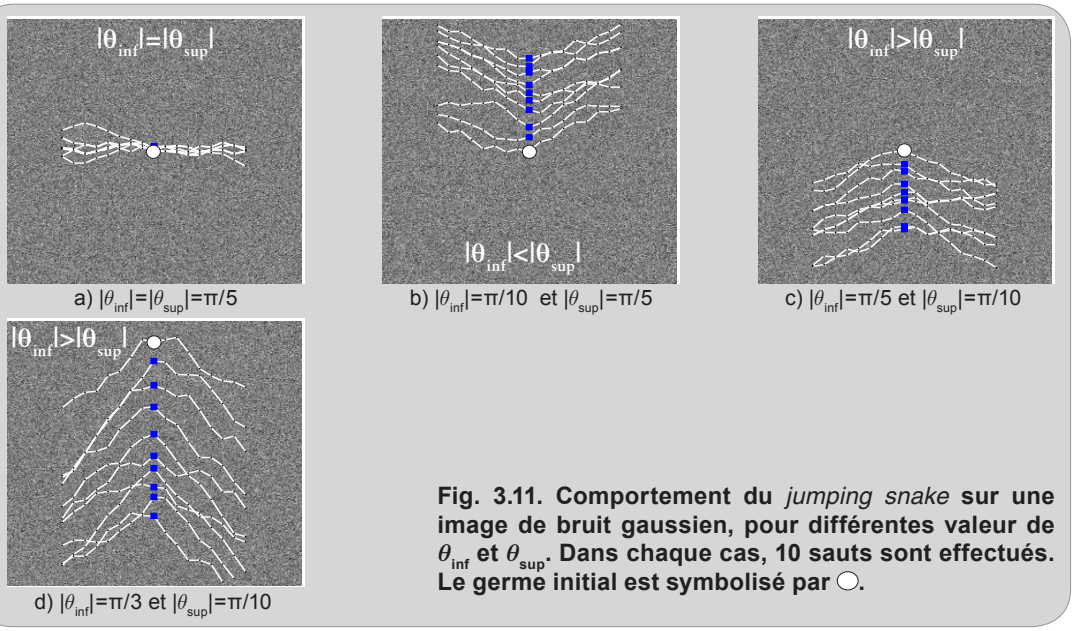
Nous verrons dans la partie 3.4 comment ces points caractéristiques permettent de construire et d'ajuster le modèle paramétrique des lèvres.

3.3.3 Influence et réglage des paramètres du jumping snake

Contrairement aux *contours actifs* classiques, le réglage des paramètres du *jumping snake* est facile et intuitif. En général, il peut être déduit directement de la taille des lèvres. De plus, il est possible de segmenter un grand nombre d'images avec un jeu unique de paramètres, la convergence étant assurée tant que quelques conditions sont remplies.

3.3.3.1 Réglage des angles de recherche θ_{inf} et θ_{sup}

S'il n'y a pas de zone de fort gradient dans le voisinage du snake, la direction globale des «branches» à droite et à gauche du germe dépend du choix de θ_{inf} et θ_{sup}, les limites angulaires de la zone de recherche. Lorsque $|\theta_{inf}|=|\theta_{sup}|$, le snake a tendance à rester globalement horizontal. La figure 3.11 présente les positions successives d'un snake après 10 sauts et 10 phases de croissance sur une image de bruit gaussien, pour différentes valeurs de θ_{inf} et θ_{sup}. Sur la figure 3.11-a, on a fixé $|\theta_{inf}|=|\theta_{sup}|=\pi/5$. On peut constater que, même après 10 sauts, le germe (symbolisé par le carré bleu au milieu du snake) a très peu bougé. De plus, les parties droite et gauche des snakes suivent globalement une direction horizontale. Si $|\theta_{inf}|<|\theta_{sup}|$, les zones de recherche supérieures sont favorisées et les branches ont tendance à «monter». Cette propriété est illustrée à la figure 3.11-b, où $|\theta_{inf}|= \pi/10$ et $|\theta_{sup}|= \pi/5$. Comme les points du snake ont plus de chance d'être situés au-dessus du germe, le saut (calculé grâce à l'équation 3.5) se fera préféren-



a) $|\theta_{inf}|=|\theta_{sup}|=\pi/5$
b) $|\theta_{inf}|=\pi/10$ et $|\theta_{sup}|=\pi/5$
c) $|\theta_{inf}|=\pi/5$ et $|\theta_{sup}|=\pi/10$
d) $|\theta_{inf}|=\pi/3$ et $|\theta_{sup}|=\pi/10$

Fig. 3.11. Comportement du *jumping snake* sur une image de bruit gaussien, pour différentes valeur de θ_{inf} et θ_{sup}. Dans chaque cas, 10 sauts sont effectués. Le germe initial est symbolisé par ○.

$[\theta_{inf}, \theta_{sup}]$	nombre de sauts	durée de la convergence
$[-2\pi/5, \pi/6]$	5	0.6 s
$[-\pi/3, \pi/6]$	5	0.4 s
$[-3\pi/10, \pi/6]$	6	0.4 s
$[-\pi/5, \pi/6]$	7	0.35 s

Tab. 3.1. Nombre de sauts et durée de la convergence pour quelques valeurs de $[\theta_{inf}, \theta_{sup}]$ et pour une résolution angulaire constante.

$[\theta_{inf}, \theta_{sup}]$	nombre de sauts	durée de la convergence
$[-2\pi/5, \pi/6]$	5	0.25 s
$[-\pi/3, \pi/6]$	5	0.25 s
$[-3\pi/10, \pi/6]$	6	0.29 s
$[-\pi/5, \pi/6]$	7	0.34 s

Tab. 3.2. Nombre de sauts et durée de la convergence pour quelques valeurs de $[\theta_{inf}, \theta_{sup}]$ et pour un nombre de candidats constant.

tiellement vers le haut. A l'opposé, si $|\theta_{inf}|>|\theta_{sup}|$, le snake aura tendance à «tomber» car la plupart de ses points seront situés sous le germe (voir figure 3.11-c). Pour notre application, le germe initial est positionné au-dessus de la bouche. Le snake doit donc descendre vers le contour supérieur des lèvres. Par conséquent, **il faut choisir $|\theta_{inf}|>|\theta_{sup}|$.**

Ensuite, suivant la valeur des angles, l'amplitude des sauts sera plus ou moins importante. La figure 3.11-d présente l'évolution d'un snake pour lequel $|\theta_{inf}|= \pi/3$ et $|\theta_{sup}|= \pi/10$. On peut constater que les sauts sont plus importants que pour le snake de la figure 3.11-c. Comme $|\theta_{inf}|$ a une valeur supérieure, la zone de recherche basse est plus étendue et les points du snake ont tendance à être plus bas. Par conséquent, les sauts ont une amplitude plus forte. L'accroissement de $|\theta_{inf}|$ permettra donc de parcourir une distance plus importante avec moins de sauts. Cependant, si l'écart angulaire entre 2 points candidats est constant, c'est-à-dire si la résolution angulaire est constante, l'accroissement de l'angle de recherche oblige à considérer un nombre de candidats plus important (voir figure 3.12). Le tableau 3.1 présente quelques résultats représentatifs obtenus pour différentes valeurs de θ_{inf} et θ_{sup}, **avec une résolution angulaire constante**. Comme on pouvait s'y attendre, le nombre de sauts pour rejoindre le contour supérieur des lèvres diminue lorsque $|\theta_{inf}|$ augmente. Cependant, on constate que la durée de la convergence augmente car le calcul de chaque point du snake est plus long. A l'opposé, si on utilise **un nombre constant de candidats**, le temps de calcul de chaque maillon du snake est constant quelle que soit la valeur de θ_{inf} (voir figure 3.12). Le tableau 3.2 présente les résultats obtenus avec les mêmes angles de recherche que le tableau 3.1. On constate que le nombre de sauts et le temps de calcul diminuent lorsque $|\theta_{inf}|$ augmente. Pour la détermination de θ_{inf} et θ_{sup}, **nous avons donc utilisé un nombre constant de candidats égal à 8**.

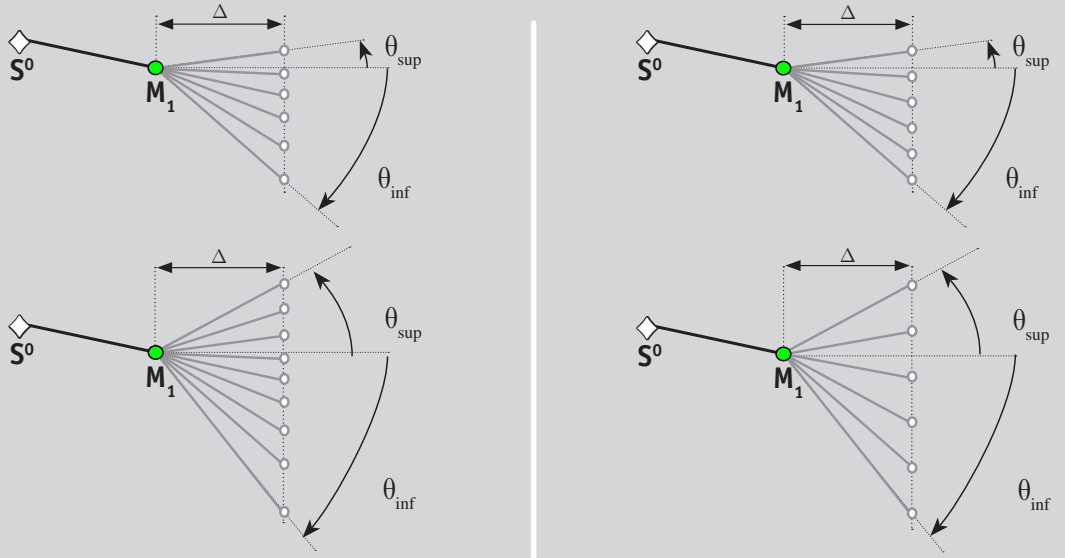


Fig. 3.12. Recherche des points du snake parmi un ensemble de candidats (○), pour une résolution angulaire constante (à gauche) et pour un nombre de candidats constant (à droite).

L'angle maximum généralement observé sur le contour supérieur des lèvres est d'environ $\pi/5$. Par conséquent, il faut que :

$$\begin{cases} |\theta_{sup}| \geq \frac{\pi}{5} \\ |\theta_{inf}| \geq \frac{\pi}{5} \end{cases} \qquad \text{(eq. 3.7)}$$

On choisit donc :

$$|\theta_{sup}| = \pi/5 \qquad \text{(eq. 3.8)}$$

A priori, il est intéressant d'attribuer une forte valeur à θ_{inf} car cela diminue le nombre de sauts nécessaires. Cependant, comme le nombre de candidats considérés est constant, un angle trop important conduit à une résolution angulaire médiocre. Dans ce cas, la position du snake final est souvent approximative. Après quelques essais, nous avons choisit :

$$|\theta_{inf}| = \pi/3 \qquad \text{(eq. 3.9)}$$

Cette valeur offre un bon compromis entre précision et rapidité.

3.3.3.2 Réglage de Δ et de N

N permet de régler le nombre de points de chaque branche du snake (voir figure 3.6). Théoriquement, il est possible de choisir N=1, ce qui conduit à un snake comportant 3 points.

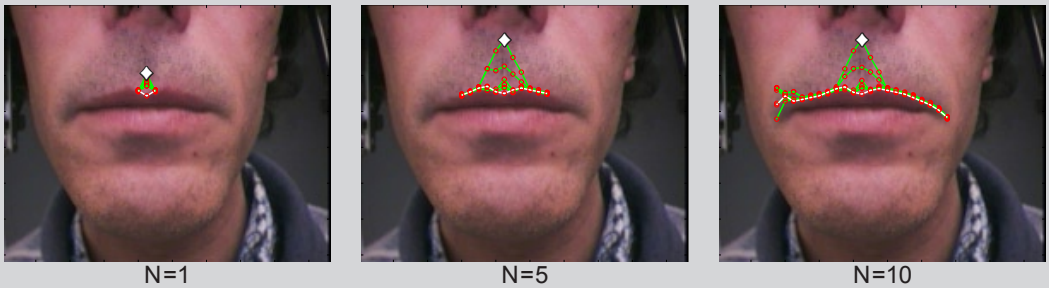


Fig. 3.13. Allure du snake pour différente valeurs de N, avec Δ constant. Le germe initial est symbolisé par ◇ et le snake final par la ligne blanche.

Cependant, dans ce cas, le contour dessiné par le snake final est trop approximatif (si Δ est grand) ou trop court (si Δ est petit) pour pouvoir estimer la position des points caractéristiques de l'arc de Cupidon (voir figure 3.13). De plus, une faible valeur de N réduit la zone de convergence du snake. A l'opposé, une forte valeur de N augmente la zone de convergence et permet d'approcher le contour des lèvres avec une bonne précision, mais alourdit également le calcul. Nous avons estimé qu'une dizaine ou une quinzaine de points régulièrement répartis sur le contour supérieur des lèvres permettait d'obtenir une estimation relativement rapide de la position de P_2, P_3 et P_4. **Nous avons donc choisi N=6**, c'est-à-dire un snake à 13 points.

Le dernier paramètre à régler est la distance horizontale entre les points, notée Δ. Il est possible d'en estimer une valeur adéquate en considérant quelques contraintes géométriques simples. Tout d'abord, il faut que le snake soit au moins aussi long que l'arc de Cupidon pour qu'on puisse déterminer la position de P_2, P_3 et P_4. Cette condition peut s'écrire :

$$2N\Delta > D_{24} \qquad \text{(eq. 3.10)}$$

où D_{24} est la distance horizontale entre P_2 et P_4. Ensuite, il faut que l'espacement entre les points soit suffisamment faible. On considère que, pour obtenir une détection précise des points caractéristiques, il doit y avoir au moins un point entre P_2 et P_4. Cette condition s'écrit :

$$2\Delta < D_{24} \qquad \text{(eq. 3.11)}$$

Dans les bases d'images que nous avons utilisées, la largeur des bouches varie de 50 à 100 pixels, ce qui correspond approximativement à une distance D_{24} comprise entre 15 et 30 pixels. Par conséquent, si on considère que N=6, les équations 3.10 et 3.11 peuvent s'écrire :

$$\begin{cases} \Delta > \dfrac{D_{24,MAXI}}{2N} \\ \Delta < \dfrac{D_{24,MINI}}{2} \end{cases} \Leftrightarrow \begin{cases} \Delta > \dfrac{30}{12} \\ \Delta < \dfrac{15}{2} \end{cases} \Leftrightarrow \quad \Delta \in [2.5, 7.5]$$

Nous avons choisi Δ=5. D'après les équations 3.10 et 3.11, cette valeur permet théoriquement de segmenter des lèvres dont la distance D_{24} varie de 10 à 60 pixels, ce qui correspond approximativement à des largeurs de bouches comprises entre 35 et 190 pixels.

Finalement, il est possible de segmenter un grand nombre d'images avec un jeu unique de paramètres $(\theta_{inf}, \theta_{sup}, N, \Delta)=(-\pi/3, \pi/5, 6, 5)$, la convergence étant assurée tant que quelques conditions sont remplies. En théorie, ces paramètres sont utilisables si la largeur de la bouche est comprise approximativement entre 35 et 190 pixels. Dans la partie suivante, nous verrons que certaines contraintes supplémentaires restreignent cet intervalle théorique.

3.4 Extraction du contour

3.4.1 Modèle polynomial

Comme il a été dit dans la partie 2.4.1.2, de nombreux modèles analytiques de bouche ont déjà été proposés dans la littérature. Le plus basique utilise 2 paraboles. Il est très facile à ajuster, mais, comme le montre la figure 3.14-b, il est trop simple pour représenter correctement les contours des lèvres. Certains auteurs ont proposé d'utiliser 2 paraboles pour le contour supérieur (figure 3.14-c) ou de remplacer les paraboles par des quartiques (figure 3.14-d). Cela améliore le précision, mais ces modèles restent limités par leur rigidité, particulièrement dans le cas de formes asymétriques.

Le modèle que nous proposons (présenté à la figure 3.14-e) est suffisamment flexible pour représenter les spécificités de lèvres très différentes. Il est composé de 5 courbes indépendantes. Chacune d'entre elles décrit une partie du contour labial. Entre P_2 et P_4, l'arc de Cupidon est dessiné par une ligne brisée passant par P_3. Les autres parties du contour sont représentées par des courbes polynomiales cubiques γ_i. De plus, chaque cubique a une dérivée nulle aux points P_2, P_4 ou P_6. Par exemple, la dérivée de γ_1 s'annule en P_2.

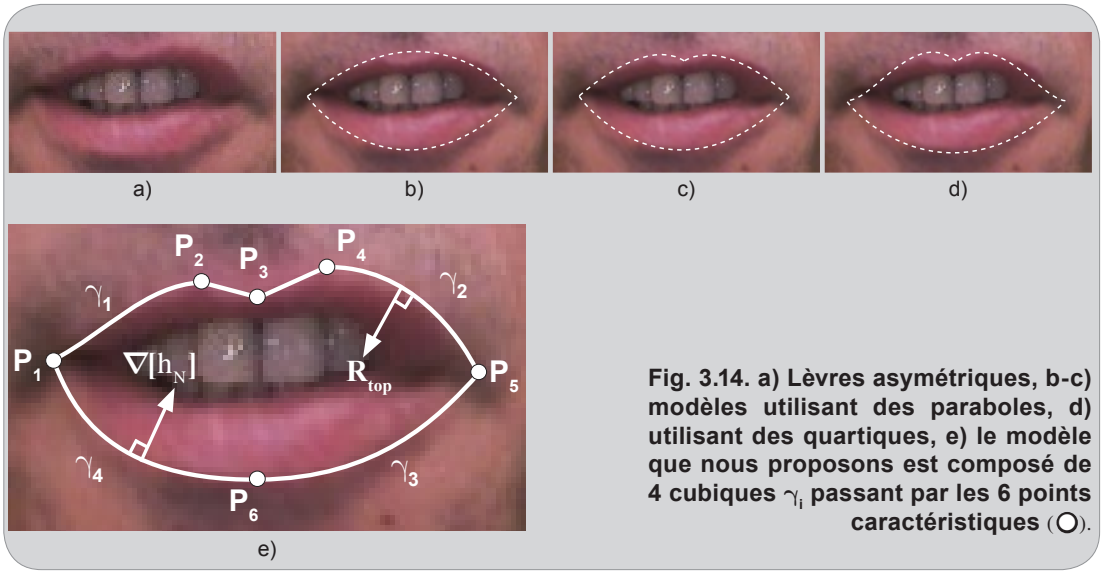


Fig. 3.14. a) Lèvres asymétriques, b-c) modèles utilisant des paraboles, d) utilisant des quartiques, e) le modèle que nous proposons est composé de 4 cubiques γ_i passant par les 6 points caractéristiques (○).

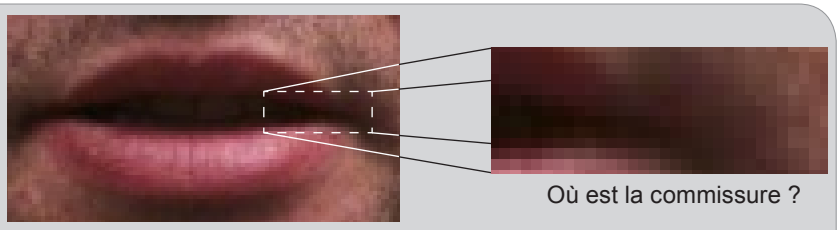


Fig. 3.15. La détermination de la position des commissures est très difficile si l'on utilise seulement un critère local.

3.4.2 Ajustement du modèle

L'ajustement du modèle et la détection des commissures de la bouche sont étroitement liés. Pour trouver les commissures, un expert humain utilise implicitement la forme globale de la bouche. Il suit les contours haut et bas, les prolonge lorsqu'ils ne sont plus suffisamment visibles, et place finalement la commissure à leur intersection. Par conséquent, les contours et les commissures sont trouvés en une seule et même opération. Sur la figure 3.15, le coin de la bouche a été extrait de l'image. On constate qu'il est difficile de localiser avec précision la commissure en utilisant seulement des informations locales, alors que l'opération ne pose pas de problème sur l'image de la bouche entière. L'algorithme que nous proposons ici exploite cette propriété.

Une courbe polynomiale cubique est définie de manière unique si ses 4 paramètres sont connus. Son expression analytique est :

$$y = ax^3+bx^2+cx+d \qquad \text{(eq. 3.12)}$$

Ici, les cubiques sont construites en utilisant 2 hypothèses :

- elles passent par P_2, P_4 et P_6
- leur dérivée s'annule en P_2, P_4 et P_6.

Ces contraintes fournissent 2 équations qui font passer le nombre de paramètres à estimer de 4 à 2 pour chaque cubique. Par conséquent, il suffit de connaître 2 points supplémentaires par cubique pour ajuster le modèle. Ces points manquants sont détectés dans les parties les plus «fiables» du contour, c'est-à-dire dans des zones proches de P_2, P_4 et P_6 (où le contour est très

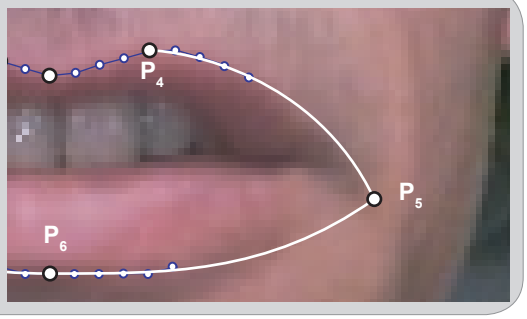


Fig. 3.16. La méthode directe et intuitive est imprécise. Quelques points additionnels (petits points) sont utilisés pour calculer les cubiques (lignes blanches). Ces courbes se rejoignent loin de la commissure réelle.

bien dessiné). Quelques points du contour supérieur ont déjà été trouvés par le *jumping snake*. En revanche, un seul point du contour inférieur est connu (P_6). Pour obtenir des points supplémentaires du contour bas, on fait croître un snake en utilisant P_6 comme germe. Les points additionnels supérieurs et inférieurs sont symbolisés par des petits points sur la figure 3.16.

Maintenant qu'il y a suffisamment de points connus sur chacune des parties du contour, il est théoriquement possible de calculer les courbes γ_i et de placer les commissures à leurs intersections. Cependant, cette méthode directe et intuitive conduit à des résultats très approximatifs. Les points utilisés pour calculer les courbes sont proches les uns des autres, et un petit déplacement de l'un d'entre eux conduit à un résultat complètement différent. Il est également possible d'utiliser plus de 2 points supplémentaires, comme le montre la figure 3.16. Dans ce cas, les courbes sont calculées en utilisant la méthode des *moindres carrés*. Utiliser 3 ou 4 points (au lieu de 2) améliore la précision, mais l'estimation des commissures reste très sensible à leur position.

La méthode directe fournit des résultats très approximatifs. Nous avons donc adopté une approche légèrement différente. Nous utilisons toujours quelques points fiables additionnels pour calculer les courbes, mais nous supposons également que les commissures (P_1 et P_5) sont connues. Cette hypothèse apporte une nouvelle équation qui fait passer le nombre de paramètres à estimer de 2 à 1 pour chaque cubique. Par conséquent, la minimisation des *moindres carrés* s'assimile à une régression linéaire et s'effectue très rapidement. De plus, les courbes résultantes sont beaucoup moins sensibles à la position des points. En d'autres termes, à une commissure donnée correspond un couple unique et facilement calculable de cubiques. L'ajustement du modèle est donc effectué en trouvant les commissures qui donnent les «meilleures courbes». Nous reviendrons sur cette notion de «meilleure courbe» un peu plus loin.

Il est évident qu'une recherche exhaustive sur toute l'image serait beaucoup trop longue. Heureusement, une hypothèse très simple permet de réduire la zone de recherche à quelques dizaines de pixels. Comme Delmas dans [Delmas, 2000], nous supposons que les commissures sont situées dans des zones sombres proches des lèvres. Un chaînage des pixels les plus sombres est donc effectué pour obtenir la ligne L_{mini} représentée en blanc sur la figure 3.18. Pour que L_{mini} soit située dans la bouche, le chaînage est initialisé sur le point le plus sombre du segment P_3P_6. La *chaîne* se propage ensuite vers la droite et la gauche en passant par les pixels les plus sombres proches de ses extrémités (voir figure 3.17). L_{mini} pourrait également être construite par une simple détection des pixels les plus sombres pour chaque colonne de l'image, comme

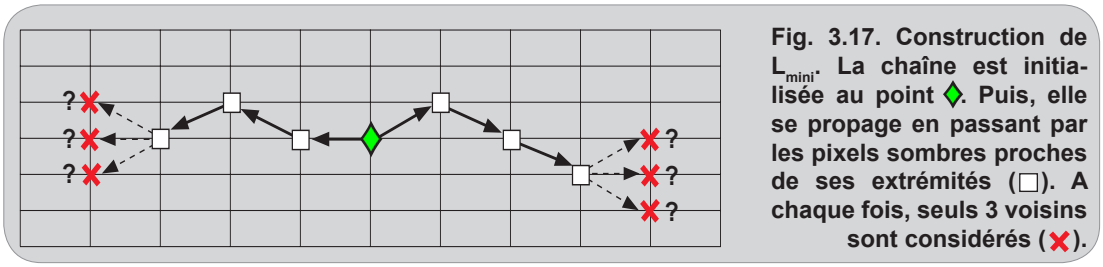


Fig. 3.17. Construction de L_{mini}. La chaîne est initialisée au point ◆. Puis, elle se propage en passant par les pixels sombres proches de ses extrémités (□). A chaque fois, seuls 3 voisins sont considérés (✕).

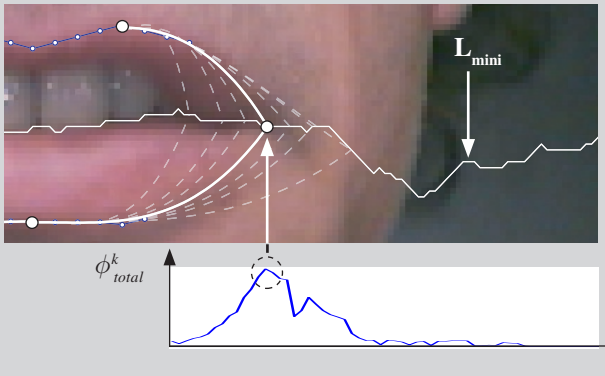


Fig. 3.18. Le maximum de ϕ^k_{total} donne la position de la commissure le long de L_{mini}. Les lignes pointillées sont les cubiques associées à quelques commissures testées sur L_{mini}.

proposé par Delmas dans [Delmas, 2000]. Cependant, la discontinuité de la ligne ainsi obtenue est très gênante pour la suite de notre algorithme.

Chaque point de la ligne L_{mini} est testé et correspond à un couple unique de cubiques supérieure et inférieure, représentées par les lignes en pointillés de la figure 3.18. Il reste donc à déterminer si les cubiques suivent correctement les lèvres. Pour cela, nous utilisons un *critère contour*. Si les cubiques supérieures γ_1 et γ_2 suivent parfaitement le contour des lèvres, elles sont orthogonales au champ de vecteurs $\boldsymbol{R}_{top}$ (voir figure 3.14). D'un autre coté, les courbes inférieures γ_3 et γ_4 doivent être orthogonales au gradient de la pseudo-teinte $\nabla[h]$. Nous calculons donc $\phi_{top,i}$ et $\phi_{low,i}$, les flux moyens à travers les courbes supérieures et inférieures respectivement :

$$\begin{cases} \phi_{top,i} = \dfrac{\int_{\gamma_i} \boldsymbol{R}_{top}.\boldsymbol{dn}}{\int_{\gamma_i} ds} \quad , \quad i \in \{1,2\} \\ \phi_{low,i} = \dfrac{\int_{\gamma_i} \nabla[h].\boldsymbol{dn}}{\int_{\gamma_i} ds} \quad , \quad i \in \{3,4\} \end{cases} \qquad \text{(eq. 3.13)}$$

où $\boldsymbol{dn}$ et ds sont respectivement le vecteur orthogonal au contour et l'abscisse curviligne. Nous considérons ensuite n positions possibles pour P_1 et P_5 le long de la ligne L_{mini}. Les courbes associées aux meilleurs candidats maximisent $\phi_{top,i}$ et minimisent $\phi_{low,i}$. Donc, de chaque coté, il faut maximiser ϕ^k_{total}, calculé comme suit :

$$\phi^k_{total} = \phi_{top,normalisé} - \phi_{low,normalisé} \quad , \quad k \in \{1,...,n\} \qquad \text{(eq. 3.14)}$$

où $\phi_{top,normalisé}$ et $\phi_{low,normalisé}$ sont les valeurs normalisées à dynamique unitaire des flux moyens :

$$\begin{cases} \phi_{top,normalisé} = \dfrac{\phi^k_{top} - \min\limits_{j\in\{1,...,n\}}(\phi^j_{top})}{\max\limits_{j\in\{1,...,n\}}(\phi^j_{top}) - \min\limits_{j\in\{1,...,n\}}(\phi^j_{top})} \\ \phi_{low,normalisé} = \dfrac{\phi^k_{low} - \min\limits_{j\in\{1,...,n\}}(\phi^j_{low})}{\max\limits_{j\in\{1,...,n\}}(\phi^j_{low}) - \min\limits_{j\in\{1,...,n\}}(\phi^j_{low})} \end{cases} \qquad \text{(eq. 3.15)}$$

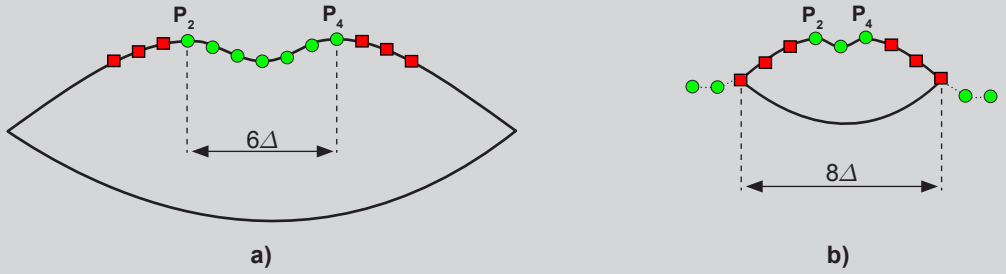


Fig. 3.19. Configurations limites. Les points supplémentaire utilisés pour le calcul des cubiques sont représentés par des carrés (■). a) largeur maximale de bouche, b) largeur minimale de bouche.

ϕ^k_{top} et ϕ^k_{low} sont les flux moyens associés à la commissure numéro k. Lorsque ϕ^k_{total} est important, la position de la commissure est fiable car les cubiques correspondantes sont proches du contour des lèvres. Comme il a été dit en introduction de cette partie, les commissures et le contour sont trouvés en une seule et même opération. La figure 3.18 montre l'évolution de ϕ^k_{total} pour différentes commissures testées, c'est-à-dire pour différentes valeurs de k. Le maximum de ϕ^k_{total} donne la position de la commissure finale le long de L_{mini}.

Dans la partie 3.3.3.2, nous avons déterminé un intervalle théorique pour les largeurs de lèvres assurant la segmentation (largeur comprise entre 35 et 190 pixels et D_{24} comprise entre 10 et 60 pixels). La méthode utilisée pour ajuster le modèle et trouver les commissures nécessite de réduire cet intervalle. En effet, pour obtenir une estimation fiable des cubiques, nous utilisons 3 points supplémentaires (les carrés ■ de la figure 3.19). Il faut donc que la bouche soit suffisamment petite pour qu'il puisse y avoir 3 points à gauche de P_2 et à droite de P_4. Par conséquent, comme le snake que nous utilisons comporte 13 points, il doit y avoir au maximum 5 points entre P_2 et P_4. La distance horizontale entre P_2 et P_4 (D_{24}) ne peut donc excéder 6Δ (c'est à dire 30 pixels), ce qui correspond à une **largeur de bouche maximale d'environ 100 pixels**. La figure 3.19-a présente cette configuration.

Ensuite, il faut que la bouche soit suffisamment grande pour que les 3 points supplémentaire utilisés soient sur le contour supérieur des lèvres. Si l'on considère que la distance D_{24} est légèrement plus petite que l'étendue horizontale d'une cubique supérieure, on peut supposer que, dans le cas limite, il n'y a qu'un seul point entre P_2 et P_4. Comme le montre la figure 3.19-b, la largeur de la bouche est alors égale à 8Δ. Nous avons choisi Δ=5, donc la **largeur théorique minimale de la bouche est de 40 pixels**.

3.5 Résultats

Les résultats de *segmentation statique* présentés dans cette section proviennent de diverses sources. Nous avons tout d'abord repris les images de la *base labiophone* utilisée par Delmas lors de sa thèse [Delmas, 2000]. Elles ont été aquises par une micro-caméra montée sur un casque fixé à la tête du locuteur. Le locuteur n'est pas maquillé et l'éclairage n'est pas uniforme. Du fait du procédé d'acquisition particulier, le cadrage est constant. La bouche est centrée et la zone du visage présente sur les images s'étend du nez au cou. Cette base contient 180 images représentant 6 locuteurs différents. Ensuite, nous avons utilisé une caméra de type *web-cam* (*Sony* EVI-D30) pour acquérir des images à cadrage variable sur une dizaine de locuteurs différents. La base ainsi constituée contient environ 2000 images. Enfin, nous avons effectué des captures d'émissions télévisées grâce à une carte d'acquisition TV de type *PCTV*.

Le grand nombre d'images à notre disposition permet de tester notre algorithme de segmentation statique dans des conditions d'éclairage et de cadrage très différentes, et sur des formes de bouches très variées. La figure 3.20 présente quelques résultats d'initialisation représentatifs. Pour chaque exemple, la position finale du *jumping snak*e ainsi que le contour extrait sont donnés. On peut tout d'abord observer que le modèle proposé permet de reproduire fidèlement de nombreuses formes de bouche. Sa grande flexibilité lui permet de s'adapter aussi bien à des lèvres «charnues et rondes» (figures 3.20-b, 3.20-c, 3.20-e) qu'à des lèvres plus fines et étirées (figures 3.20-i, 3.20-j, 3.20-l) ou asymétriques (figures 3.20-f). De plus, la méthode fonctionne quelle que soit l'ouverture de la bouche. Enfin, il est interessant de noter que la segmentation peut être effectuée même en présence de barbe (figures 3.20-c, 3.20-k) ou d'ombres très marquées (figures 3.20-g, 3.20-h).

Les zones grisées sur les images de la figure 3.20 représentent les zones d'initialisation admissibles du *jumping snake*, c'est-à-dire les zones dans lesquelles doivent se situer les germes initiaux pour permettre une convergence du contour actif sur le contour supérieur des lèvres et une détection des points caractéristiques. Comme les points P_2 et P_4 sont situés de part et d'autre du germe initial, les limites verticales des zones de convergence sont données par les limites de l'arc de Cupidon. Ensuite, le germe initial doit être situé au-dessus du contour supérieur des lèvres car, comme nous avons choisi $|\theta_{sup}|<|\theta_{inf}|$ (voir partie 3.3.3.1 et figure 3.11), le *jumping snake* a tendance à descendre. Enfin, la limite horizontale supérieure de la zone de convergence correspond au bas du nez. En effet, si le *jumping snake* passe sur le nez, il risque de s'y arrêter car les contours intérieurs des narines sont des zones pour lesquelles le gradient hybride $\boldsymbol{R}_{top}$ est fort. Le germe initial doit donc être positionné plus bas que les narines. Finalement, la zone de convergence dépend de la morphologie du locuteur. En général, elle a une étendue verticale environ égale au tiers de la largeur de la bouche et une étendue horizontale égale à D_{24}.

Dans la section précédente, nous avons montré que la segmentation peut être effectuée correctement tant que la distance D_{24} est comprise entre 10 et 30 pixels, ce qui correspond à une largeur de bouche allant de 40 à 100 pixels environ. Dans nos essais, les meilleurs résultats ont

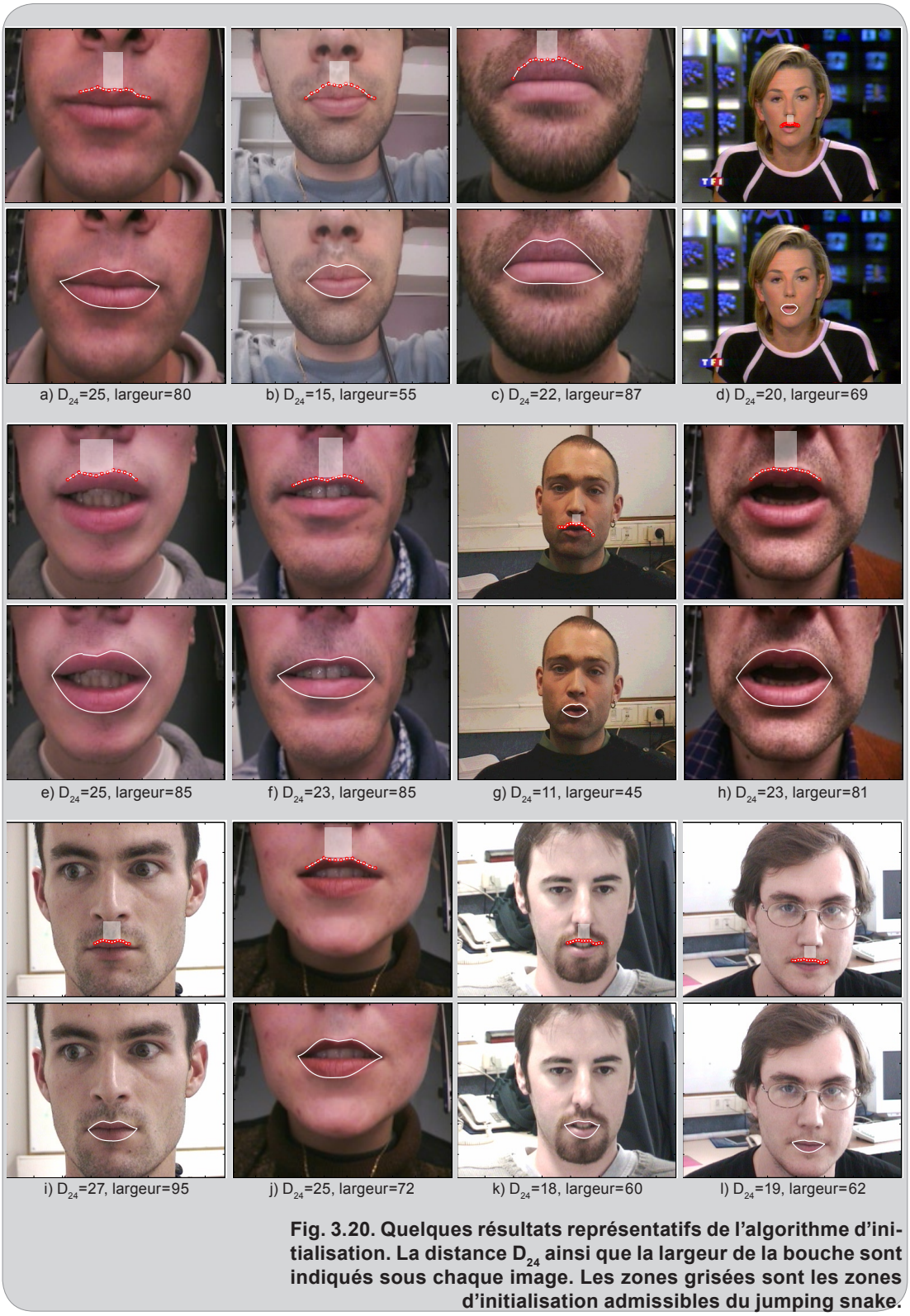


Fig. 3.20. Quelques résultats représentatifs de l'algorithme d'initialisation. La distance D_{24} ainsi que la largeur de la bouche sont indiqués sous chaque image. Les zones grisées sont les zones d'initialisation admissibles du jumping snake.

été obtenus pour des largeurs de bouche comprises entre 70 et 95 pixels. Dans cet intervalle, et pour des «apparences moyennes» (cette notion est détaillée un peu plus bas), les taux de réussite sont proches de 100%. Il est important de noter que les taux de réussite dont nous parlons sont «subjectifs» car la qualité des segmentations a été estimée «à l'oeil». Lorsque la largeur de la bouche est plus faible, les résultats se dégradent car l'incertitude sur la position des points du *jumping snake* s'accroit. Pour des largeurs de bouche comprises entre 45 et 70 pixels, les taux de réussite sont très variables selon les séquences. Cependant, toujours pour des «apparences moyennes», les taux de réussites sont au moins égaux à 50 %. Pour des largeurs inférieures à 45 pixels, les résultats sont en général mauvais car l'écartement des points du *jumping snake* est trop important pour permettre une détection fiable des points caractéristiques. De plus, comme il a été montré dans la partie 3.4.2, les commissures ne peuvent pas être correctement localisées si la largeur de la bouche est inférieure à 40 pixels. La figure 3.21-a présente un essai de segmentation d'une bouche de 28 pixels de largeur avec Δ=5. L'estimation de la position des commissures est fausse car la zone de recherche commence trop loin de la bouche. Si Δ diminue, les points du snake se resserrent et les points utilisés pour le calcul des cubiques sont sur le contour des lèvres. Sur la figure 3.21-b, la diminution de Δ permet d'effectuer une segmentation correcte des lèvres.

Même lorsque la largeur de la bouche est comprise entre 70 et 95 pixels, il arrive que, dans certaines configurations particulières, l'estimation des commisssures soit imprécise. Par exemple, si la bouche est trop étirée, la position estimée sera en général trop à l'intérieur. La figure 3.22 présente des résultats de segmentation sur des lèvres dont les côtés sont très fins. Bien que les points du snake soient correctement positionnés le long du contour supérieur, les commissures estimées sont trop proches du centre de la bouche et les bords droit et/ou gauche des lèvres sont tronqués. Cet artefact peut s'expliquer en remarquant que, dans le cas d'une bouche étirée, les zones de commissures sont des lignes sombres et ne contiennent pas la «couleur lèvres». Dès lors, elles ne sont pas incluses dans les cubiques du contour lors des maximisations

a) Δ=5 b) Δ=3

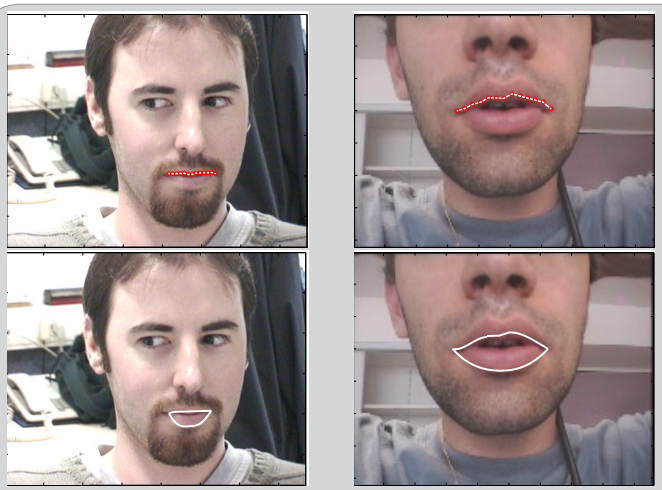
Fig. 3.21. Essais de segmentation d'une bouche de 28 pixels de largeur. a) Δ=5, b) Δ=3.

Fig. 3.22. Lorsque les côtés de la bouche sont trop fins, l'estimation des commissures peut être imprécise.

des flux moyens.

Tout comme pour les commissures, certaines configurations conduisent à une mauvaise représentation du contour supérieur. Par exemple, si le contour est trop «plat», l'arc de Cupidon n'apparait pas et la détection des points supérieurs (P_2, P_3 et P_4) est difficile. De même, si le contour est peu marqué, la position des points du snake est approximative, ce qui conduit à une estimation hasardeuse des points caractéristiques. Un tel cas de figure est présenté à la figure 3.23-a. Une bouche trop arrondie, c'est à dire la configuration inverse du cas précédent, empêche également d'effectuer une segmentation correcte. Dans ce cas, la pente du contour est trop importante pour que le snake puisse le suivre (voir figure 3.23-b).

Enfin, des mauvais postionnements du point bas ont également été relevés. Ils ont trois causes principales. Tout d'abord, s'il y a trop peu de différence chromatique entre la lèvre du bas et la peau, le contour inférieur est mal défini et la norme de $\nabla[h]$ est faible. Dans ce cas, Le minima de $\nabla_y[h]$ ne correspond pas à la limite inférieure de la bouche (voir section 3.3.2). La



a) b)

Fig. 3.23. Mauvaises représentations du contour supérieur. a) contour «plat» et peu marqué, b) contour trop pentu.

a) b) c)

Fig.3.24. mauvaise détection du point bas. a) contour inférieur mal dessiné, b) lèvre inférieure très brillante, c) décalage horizontal de la lèvre inférieure trop important.

figure 3.24-a présente une telle configuration. De même, des lèvres très brillantes conduisent également à une mauvaise estimation du minima de $\nabla_y[h]$ et à un positionnement du point bas sur la lèvre inférieure, à la limite du reflet (voir figure 3.24-b). Enfin, si le décalage horizontal entre les lèvres supérieure et inférieure est trop important, la détection précise du point bas peut être difficile. En effet, comme le contour inférieur n'est pas horizontal sous P_3, la composante $\nabla_y[h]$ du gradient n'est pas forcément importante à la frontière des lèvres, même si le contour est bien marqué. La figure 3.24-c présente une telle configuration.

3.6 Conclusion

Dans ce chapitre, nous avons tout d'abord proposé d'utiliser la pseudo-teinte car elle est relativement insensible aux variations d'illumination et permet de bien faire ressortir les lèvres par rapport à la peau. Ensuite, nous avons introduit un nouveau type de contour actif pour localiser la bouche : le *jumping snake*. Sa convergence étant «discontinue» (succession de phases de croissance et de sauts), il peut franchir plus facilement les zones bruitées de l'image et peut donc être initialisé relativement loin du contour final. La segmentation proprement dite est effectuée par un modèle de lèvres analytique composé de courbes cubiques. Bien que sa description soit compacte (il est défini par 10 paramètres), il est suffisamment flexible pour représenter des formes de lèvres très différentes. Son ajustement aux contours se fait en deux étapes : **estimation de la position** et **estimation de la forme**. Dans un premier temps, la position est obtenue à l'aide de quatre points caractéristiques situés sur le contour labial (trois points supérieurs et un point inférieur). Cette première étape permet donc d'estimer huit paramètres (i.e. quatre couples de coordonnées). L'étape suivante permet d'obtenir la forme optimale par la détermination conjoite de la position des commissures et des cubiques associées. Cette optimisation séquentielle des paramètres permet de réduire considérablement la complexité du problème. Au lieu d'estimer simultanément tous les paramètres, comme par exemple dans [Liew *et al.*, 2000], la méthode que nous proposons permet de les obtenir les uns après les autres (ou

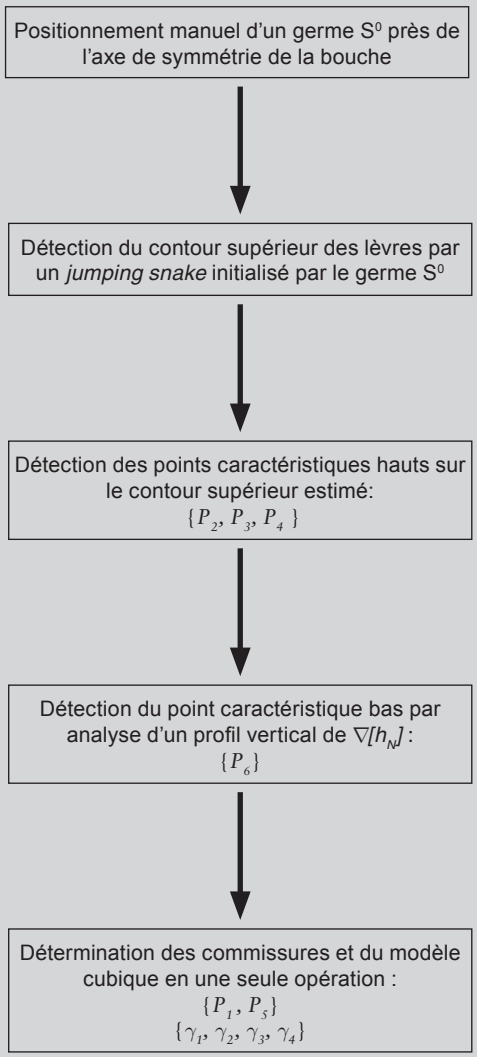


Fig. 3.25. Résumé de notre algorithme de *segmentation statique*.

bien par couple, dans le cas de la détection d'un point caractéristique).

Finalement, l'algorithme de *segmentation statique* que nous proposons (résumé à la figure 3.25) fonctionne correctement à condition que quelques hypothèses soient satisfaites. Tout d'abord, le jeu de paramètres utilisé pour faire converger le *jumping snake* impose des limites géométriques théoriques. Pour qu'une bouche puisse être segmentée, sa largeur doit être comprise entre 40 et 100 pixels. Bien que les meilleurs taux de réussite soient obtenus pour des largeurs comprises entre 70 et 95 pixels, il est possible de segmenter des bouches de largeurs comprises entre 40 et 70 pixels. Dans ce cas, les taux de réussite sont moins bons, mais restent toujours supérieurs à 50% pour des «apparences moyennes» de lèvres. Les «apparences moyennes» correspondent à la plupart des lèvres observées au repos et sans expression forcée. Elles excluent donc les formes transitoires du language et les grimaces (comme sur la figure 3.24-c),

les étirements et les protrusions très marqués (i.e. les sourires et les moues). Elles excluent également les lèvres ayant peu de différence chromatique avec la peau.

Nous verrons dans le chapitre suivant que la prise en compte des informations temporelles permet d'améliorer de manière très significative les performances de notre système de segmentation. Nous montrerons notamment que le suivi temporel permet d'effectuer la segmentation dans la plupart des cas problématiques décrits ci-dessus, pour peu que ces cas ne se présentent pas à la première image.

CHAPITRE 4

Segmentation dynamique

4.1 Introduction

Comme nous l'avons mentionné en introduction du chapitre 3, la méthode de segmentation que nous proposons comporte 2 étapes principales : l'*initialisation* et le *suivi*. Dans l'étape d'*initialisation*, le contour était détecté en utilisant une «segmentation statique». Lors du *suivi*, les résultats obtenus dans les images précédentes fournissent des informations supplémentaires susceptibles de rendre la segmentation plus robuste et plus rapide. Du fait de l'utilisation de ces informations temporelles, on parlera ici de «segmentation dynamique».

Tout d'abord, dans la partie 4.2, nous montrons comment l'algorithme de Lucas-Kanade permet de suivre les points caractéristiques d'une image à l'autre. Comme cette méthode n'utilise que le voisinage des points, elle apporte un gain de temps significatif par rapport à la technique d'extraction directe du chapitre précédent. Toutefois, nous montrons que l'accumulation des erreurs de suivi est inévitable et conduit, après quelques images, à des résultats approximatifs.

La partie 4.3 présente un algorithme original permettant de compenser ces erreurs. Dans un premier temps, les points hauts et bas sont recalés en utilisant une version simplifiée des *contours actifs*. Puis, la position des commissures est ajustée en déformant les courbes du modèle obtenu à l'image précédente.

Dans la partie 4.4, les contours finaux sont extraits. Pour cela, la forme de la bouche dans l'image précédente ainsi que les points caractéristiques sont utilisés pour calculer les courbes optimales constituant le modèle.

Enfin, dans la partie 4.5, nous commentons quelques résultats représentatifs et nous analysons les forces et les faiblesses de la méthode.

4.2 Le suivi de points

Incontestablement, l'estimation de mouvement est un problème fondamental en traitement d'image appliqué à des séquences vidéo. Dans ce domaine, de très nombreuses méthodes ont été (et continuent d'être) proposées. On peut distinguer trois approches principales.

Tout d'abord, les **méthodes différentielles** (*differential methods*) s'appuient sur l'équation de contrainte du mouvement apparent issue d' un développement de Taylor de l'équation 4.1. Parmi les différentes variantes proposées, certaines sont basées sur des dérivées du premier ordre avec ou sans contrainte de régularisation sur le champ de vecteurs vitesse [Lucas, 1984][Horn and Schunck, 1981]. Il est également possible d'utiliser des dérivées d'ordre supérieur. De plus, il est possible de réduire la sensibilité des calculs numériques en utilisant des contraintes de régularisation locales [Uras *et al.*, 1988] ou globales [Nagel, 1987].

Ensuite, les **méthodes de mise en correspondance** (*block-matching techniques*) tentent d'estimer le mouvement d'une région de l'image courante en minimisant la distance avec une region candidate de l'image suivante. En général, cette mesure de similarité est obtenue par une somme des différences inter-pixels au carré (*Sum-Squared Difference - SSD*). Comme il est évident qu'un test exhaustif de toutes les régions possibles est très coûteux en temps de calcul, de nombreux algorithmes «rapides» ont été proposés. Anandan préconise une approche de type multi-résolution en utilisant une décomposition pyramidale de l'image [Anandan, 1989]. Le déplacement est estimé itérativement en commençant par le niveau de résolution le plus grossier. Dans [Koga, 1981], Koga propose une méthode de recherche rapide (*logarithmic search*) permettant d'estimer le déplacement d'un bloc en suivant la direction de moindre déformation. Dans le même esprit, on peut également citer la technique de recherche en trois étapes utilisée dans le codeur vidéo H.263 [ITU, 1995].

Enfin, les **méthodes fréquentielles** utilisent des bancs de filtres passe-bande permettant de décomposer le signal d'entrée selon l'échelle, la vitesse et l'orientation. Dans [Heeger, 1988], Heeger analyse l'énergie à la sortie de filtres de Gabor pour estimer les vitesses. Dans [Fleet and Jepson, 1990], Fleet et Jepson préconisent d'utiliser plutot la phase des signaux de sortie car elle est beaucoup plus stable que l'amplitude.

Pour notre application, le suivi doit tout d'abord être précis. Nous avons donc écarté les *méthode de mise en correspondance* car elles ne permettent d'estimer que des déplacements entiers (ou demi-entiers pour les techniques mises en oeuvre dans les codeurs MPEG-1). De plus, les algorithmes de recherche rapide qu'elles utilisent conduisent fréquemment à des minima locaux. D'après l'étude très détaillée menée dans [Baron *et al.*, 1994], les techniques les plus fiables et les plus précises sont la *méthode différentielle* du premier ordre de Lucas et Kanade et la *méthode de phase* de Fleet et Jepson. Cependant, il est à noter que la méthode proposée par Fleet et Jepson est beaucoup plus lente car elle nécessite un grand nombre de filtrages. Finalement, comme notre algorithme doit être rapide, nous avons donc opté pour la méthode de Lucas et Kanade dont le principe général est exposé dans la partie suivante.

4.2.1 L'algorithme de Lucas-kanade

La méthode d'estimation de mouvement que nous utilisons est basée sur l'algorithme de flux optique développé par Lucas et Kanade dans [Lucas, 1984]. Dans cette méthode, on suppose que le voisinage du point suivi dans l'image I_t se retrouve dans l'image suivante I_{t+1} par une translation :

$$I_t(\boldsymbol{x}-\boldsymbol{d}(\boldsymbol{x})) = I_{t+1}(\boldsymbol{x}) \quad \text{(eq. 4.1)}$$

où $\boldsymbol{d}(\boldsymbol{x})$ est le vecteur déplacement du pixel de coordonnée $\boldsymbol{x}$ ($\boldsymbol{x}$ est une vecteur). La figure 4.1 illustre cette égalité dans le cas d'un signal mono-dimensionnel.

Considérons un voisinage R de taille $n \times n$ dans l'image de référence prise au temps t. Le but est de retrouver dans l'image suivante la région la plus ressemblante à R. On note $I_t(\boldsymbol{x})$ et $I_{t+1}(\boldsymbol{x})$ les valeurs des niveaux de gris dans ces 2 images. Pour cela, il faut minimiser une fonction coût égale à la somme des différences inter-pixels au carré :

$$\varepsilon(\boldsymbol{d}(\boldsymbol{x})) = \sum_{\boldsymbol{x}\in R} [I_t(\boldsymbol{x}-\boldsymbol{d}(\boldsymbol{x})) - I_{t+1}(\boldsymbol{x})]^2 w(\boldsymbol{x}) \quad \text{(eq. 4.2)}$$

où $w(\boldsymbol{x})$ est une fonction de pondération. En général, $w(\boldsymbol{x})$ est constante et vaut 1. Mais elle peut également prendre une forme gaussienne si on veut donner plus d'importance au centre de la fenêtre. La minimisation de la fonction ε est réalisée de manière itérative. On note $\boldsymbol{d}^i(\boldsymbol{x})$ la valeur du déplacement total calculée au début de l'itération i. Le déplacement final $\boldsymbol{d}^{i+1}(\boldsymbol{x})$ peut alors s'exprimer de la manière suivante :

$$\boldsymbol{d}^{i+1}(\boldsymbol{x}) = \boldsymbol{d}^i(\boldsymbol{x}) + \boldsymbol{\Delta d}^i(\boldsymbol{x}) \quad \text{(eq. 4.3)}$$

où $\boldsymbol{\Delta d}^i(\boldsymbol{x})$ est le déplacement incrémental à déterminer avec une précision sub-pixel. Dans toute la suite de cette section, on suppose que le voisinage considéré ne subit pas de déformation. Par conséquent, la valeur du déplacement est la même pour tous les pixels de R. L'équation précédente peut donc être écrite plus simplement de la manière suivante :

$$\boldsymbol{d}^{i+1} = \boldsymbol{d}^i + \boldsymbol{\Delta d}^i \quad \text{(eq. 4.4)}$$

Ainsi, on peut écrire :

$$I_t(\boldsymbol{x}-\boldsymbol{d}^{i+1}) = I_t(\boldsymbol{x}-(\boldsymbol{d}^i + \boldsymbol{\Delta d}^i)) \quad \text{(eq. 4.5)}$$

En utilisant un développement de Taylor au premier ordre, cette équation devient :

$$I_t(\boldsymbol{x}-\boldsymbol{d}^{i+1}) \approx I_t(\boldsymbol{x}-\boldsymbol{d}^i) - \boldsymbol{g}^T \boldsymbol{\Delta d}^i \quad \text{(eq. 4.6)}$$

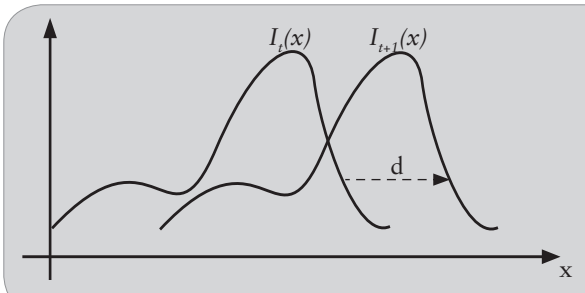


Fig. 4.1. Un voisinage dans l'image I_t peut être retrouvé dans l'image I_{t+1} par une translation de vecteur d.

où $\boldsymbol{g}$ est le vecteur gradient :

$$\boldsymbol{g} = \begin{pmatrix} \left(\frac{\partial I_t}{\partial x}\right)_{\boldsymbol{x}-\boldsymbol{d}^i} \\ \left(\frac{\partial I_t}{\partial y}\right)_{\boldsymbol{x}-\boldsymbol{d}^i} \end{pmatrix} \qquad \text{(eq. 4.7)}$$

En tenant compte de cette linéarisation, l'expression de la fonction coût ε (équation 4.2) s'écrit :

$$\begin{aligned} \varepsilon(\boldsymbol{d}) &\approx \sum_{\boldsymbol{x}\in R} [\, I_t(\boldsymbol{x}-\boldsymbol{d}^i) - \boldsymbol{g}^T \boldsymbol{\Delta d}^i - I_{t+1}(\boldsymbol{x})\,]^2 w(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}\in R} [\, h - \boldsymbol{g}^T \boldsymbol{\Delta d}^i\,]^2 w(\boldsymbol{x}) \end{aligned} \qquad \text{(eq. 4.8)}$$

où $h = I_t(\boldsymbol{x}-\boldsymbol{d}^i) - I_{t+1}(\boldsymbol{x})$.

Il s'agit d'obtenir la valeur de $\boldsymbol{\Delta d}^i$ qui minimise ε. On dérive donc $\varepsilon(\boldsymbol{d})$ par rapport à $\boldsymbol{\Delta d}^i$:

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\Delta d}^i} = 2 \sum_{\boldsymbol{x}\in R} (\, h - \boldsymbol{g}^T \boldsymbol{\Delta d}^i\,)\, \boldsymbol{g}\, w(\boldsymbol{x}) \qquad \text{(eq. 4.9)}$$

Or, $(\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Delta d}^i)\boldsymbol{g} = (\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})\boldsymbol{\Delta d}^i$. Donc, l'équation 4.9 peut s'écrire :

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\Delta d}^i} = 2\left(\sum_{\boldsymbol{x}\in R} h\boldsymbol{g}\, w(\boldsymbol{x})\right) - 2\left(\sum_{\boldsymbol{x}\in R} \boldsymbol{g}\boldsymbol{g}^T w(\boldsymbol{x})\right)\boldsymbol{\Delta d}^i \qquad \text{(eq. 4.10)}$$

L'annulation de cette dérivée conduit à l'égalité suivante :

$$\boxed{G\boldsymbol{\Delta d}^i = \boldsymbol{e}} \qquad \text{(eq. 4.11)}$$

avec :

$$\begin{cases} G = \sum_{\boldsymbol{x}\in R} \boldsymbol{g}\boldsymbol{g}^T w(\boldsymbol{x}) \\ \boldsymbol{e} = \sum_{\boldsymbol{x}\in R} (\, I_t(\boldsymbol{x}-\boldsymbol{d}^i) - I_{t+1}(\boldsymbol{x}))\, \boldsymbol{g} w(\boldsymbol{x}) \end{cases} \qquad \text{(eq. 4.12)}$$

L'équation 4.11 est la relation fondamentale de l'algorithme de Lucas-Kanade. Pour tout couple d'images adjacentes, la matrice G peut être calculée à partir de la première image en calculant le gradient de la luminance sur le voisinage du point à suivre. D'autre part, le vecteur $\boldsymbol{e}$ est obtenu en multipliant ce gradient par la différence entre les 2 fenêtres d'observation. Finalement, le déplacement incrémental recherché $\boldsymbol{\Delta d}^i$ est la solution du système 4.11.

Au début du processus, on fixe $\boldsymbol{d}^0 = [0\ 0]^{\mathrm{T}}$. Puis, le déplacement total est calculé en plusieurs itérations. A chaque fois, l'équation 4.11 permet de déterminer le déplacement incrémental à effectuer. La figure 4.2 illustre le déroulement de l'algorithme en présentant les positions successives d'un point suivi. Lorsque le déplacement calculé $\boldsymbol{\Delta d}^i$ devient inférieur à un seuil ou que le nombre d'itérations dépasse une certaine limite, le processus s'arrête.

Fig. 4.2. A partir de la position d'un point dans l'image à l'instant t (○), l'algorithme de Lucas-Kanade estime la position correspondante dans l'image suivante (□). Les positions intermédiaires calculées à chaque itération sont symbolisées par des petits points blancs.

4.2.2 Application aux lèvres et réglage des paramètres

L'algorithme de Lucas-Kanade est paramétré par trois coefficients : les deux paramètres d'arrêt (le nombre maximal d'itérations et le seuil d'arrêt sur le déplacement incrémental) et la taille de la fenêtre d'observation (n). Les 2 premiers paramètres sont relativement faciles à régler. Le graphique de la figure 4.3 présente l'évolution moyenne du déplacement incrémental Δd^i (calculée sur toute notre base d'image et pour un suivi des 6 points caractéristiques) en fonction du nombre d'itérations. Cette évolution moyenne a été calculée pour différentes tailles de fenêtre n. On peut constater que la plus grande partie des déplacements s'effectue lors des toutes premières itérations et que leur amplitude diminue ensuite très rapidement. Pour le seuil d'arrêt sur Δd^i, nous avons fixé une valeur de 0.1 pixel puisque, au-delà de cette limite, l'amplitude des sauts devient négligeable face au déplacement total. Ce seuil est symbolisé par la

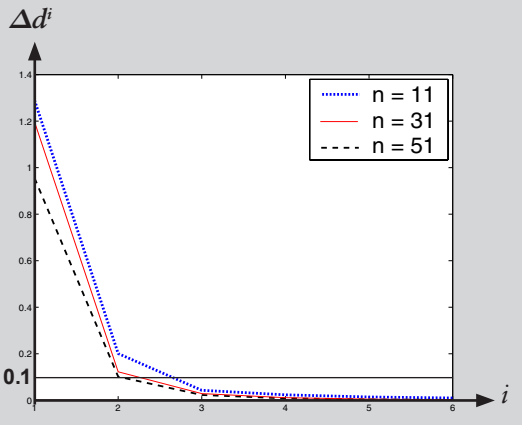


Fig. 4.3. Evolution de Δd^i en fonction de i, pour différentes tailles de fenêtres n.

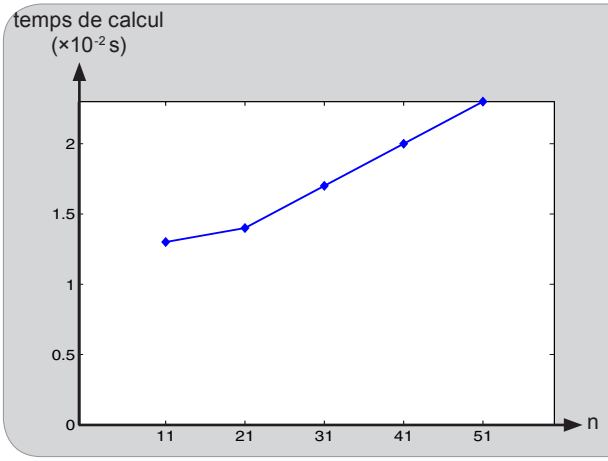


Fig. 4.4. Evolution du temps de calcul moyen pour le suivi d'un point, en fonction de la taille de la fenêtre.

ligne horizontale sur la figure 4.3. En examinant ce graphique, nous avons également fixé un nombre maximal d'itérations égal à 6. Quelle que soit la taille de la fenêtre, en moyenne les déplacements sont très faibles à la sixième itération. Si ce n'est pas le cas lors du suivi d'un point particulier, l'algorithme est probablement en train de «se perdre» et il faut donc l'arrêter.

Le réglage de la taille n de la fenêtre est un plus délicat. Le premier aspect de ce problème est la complexité calculatoire du suivi. En effet, si l'on choisit un petit voisinage, le calcul sera rapide car l'intégration se fait sur une fenêtre réduite. A l'opposé, une forte valeur de n conduit à un calcul plus lent. Le graphique de la figure 4.4 présente l'évolution du temps de calcul moyen pour le suivi d'un point en fonction de la taille de la fenêtre. Ce temps augmente en même temps que n. Cependant, alors qu'on attendait une évolution quadratique en n^2 (car le nombre de points du voisinage à traiter augmente avec le carré de n), on obtient une relation plutôt linéaire. Cela est probablement dû au fait que, lorsqu'on utilise une petite fenêtre, un nombre plus important d'itérations est nécessaire. En effet, sur le graphique de la figure 4.3, la courbe correspondant à n=11 franchit le seuil d'arrêt après celle correspondant à n=51.

Ensuite, au-delà du temps de calcul, le paramètre n a une influence directe sur la qualité du suivi effectué. L'utilisation d'une petite fenêtre conduit à un calcul rapide, mais les résultats obtenus sont peu fiables car le nombre de pixels considérés est faible. Le phénomène est d'autant plus marqué que le point à suivre est situé dans des zones d'occultation ou d'apparition. C'est le cas, par exemple, des commissures qui sont en général très proches de l'intérieur de la bouche (où la langue, les dents et des zones sombres peuvent apparaître et disparaître très rapidement). La figure 4.5 présente des résultats de suivi des 6 points caractéristiques sur la séquence *Nico* pour différentes tailles de fenêtre d'observation. On constate qu'avec n=11, le suivi est approximatif (voire complètement faux pour la commissure gauche). Lorsque n augmente, on peut remarquer que le suivi est meilleur. Cette observation qualitative est confirmée par les mesures du tableau 4.1 qui fait apparaître les écarts moyens par rapport à un étiquetage manuel sur la séquence *Nico*. Les points supérieurs (P_2, P_3 et P_4) sont les mieux suivis car ils sont situés dans des zones peu déformables et aux contours relativement bien marqués. A l'opposé, les erreurs de

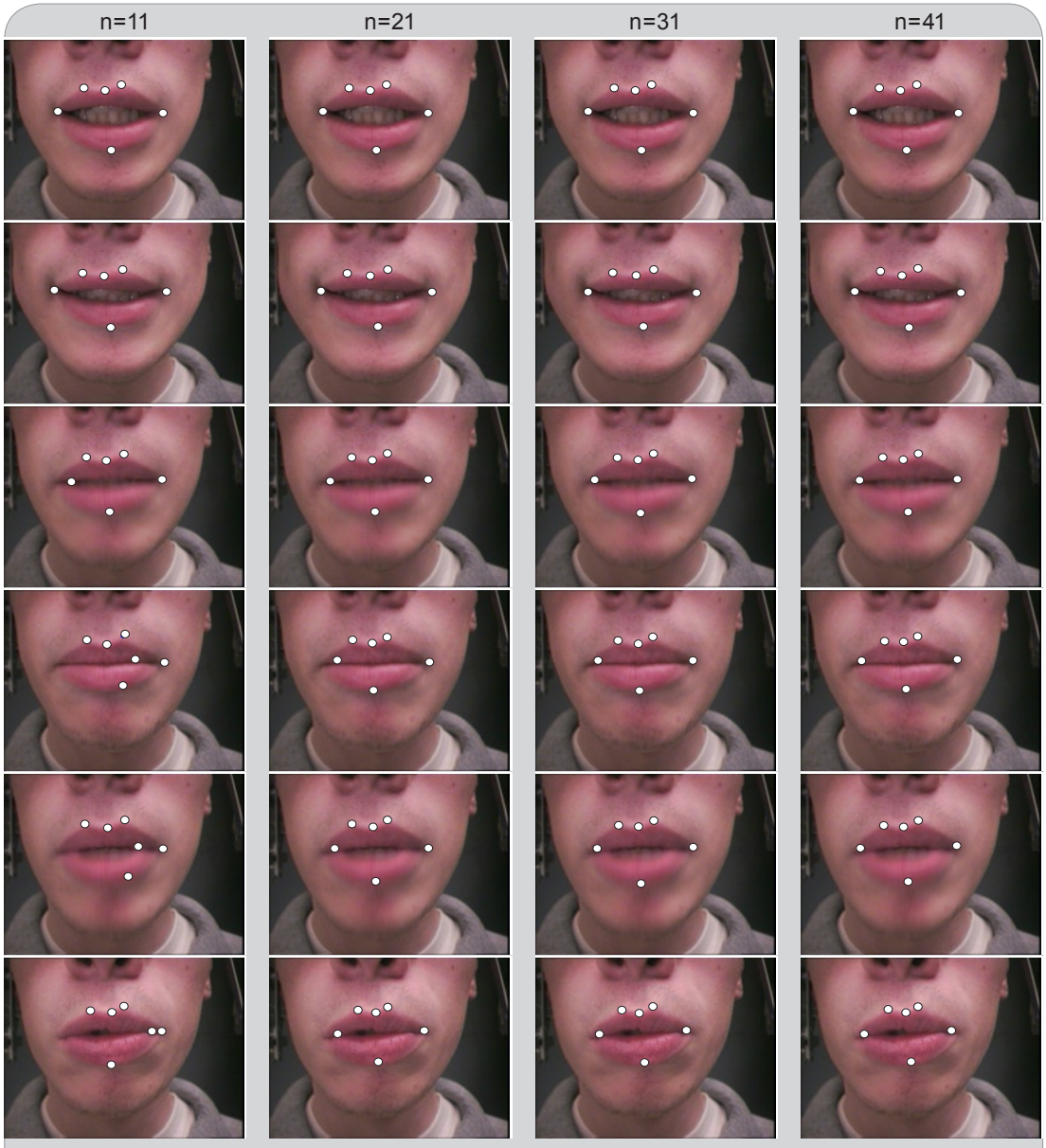


Fig. 4.5. Résultats du suivi des 6 points caractéristiques sur la séquence *Nico* pour différentes tailles de fenêtre.

n	P_1	P_2	P_3	P_4	P_5	P_6
11	58%	4.3%	4.3%	3.5%	6.9%	7.3%
21	10%	3.9%	4.3%	3.1%	6.9%	7%
31	8.5%	3.9%	4.4%	3.1%	6.4%	6.3%
41	8.5%	3.8%	4.6%	3.5%	4.7%	6.9%

Tab. 4.1. Ecarts moyens des points suivis sur la séquence *Nico* par rapport aux points obtenus par un étiquetage manuel. Les écarts sont donnés en pourcentage de la largeur de la bouche.

suivi sur les commissures (P_1 et P_5) sont relativement importantes, même pour des fortes valeurs de n. Ce phénomène a 2 origines. Tout d'abord, les commissures sont situées dans des zones aux contours peu marqués. Or, l'algorithme de suivi de Lucas-Kanade fonctionne d'autant mieux qu'il est appliqué à des points situés sur des contours forts. Ensuite, les fenêtres associées aux commissures empiètent généralement sur l'intérieur de la bouche qui est une zone d'occultation et d'apparition. Par conséquent, comme les mouvements articulatoires peuvent être rapides, les voisinages des commissures dans 2 images consécutives peuvent être très différents. Cet empiètement des fenêtres sur l'intérieur de la bouche peut également apparaître avec les autres points si n est trop grand. Cela explique que, dans le tableau 4.1, les erreurs de suivi des points P_3 et P_6 sont plus importantes avec $n=41$ qu'avec $n=31$.

Les erreurs de suivi sur le point bas (P_6) sont également importantes, quelle que soit la valeur de n. Ce point est situé sur un contour qui est généralement horizontal. Dans ce cas, la première composante du vecteur $\boldsymbol{g}$ (voir équation 4.7) est faible et l'estimation du mouvement horizontal (obtenu par l'équation 4.11) est peu fiable. On peut d'ailleurs observer sur la figure 4.5 que seule la position horizontale du point P_6 est approximative. Ce phénomène très classique est connu sous le nom de *problème d'ouverture* [Hildreth, 1984].

Pour rendre le suivi des commissures plus robuste, nous avons décalé leur fenêtre de manière à ce qu'elle n'empiète pas sur l'intérieur de la bouche, comme indiqué à la figure 4.6. Ainsi, les variations brusques de l'apparence de l'intérieur de la bouche ne sont pas *vues* dans les fenêtres d'observation. La figure 4.7 présente les résultats de suivi sur la séquence *Nico* , obtenus avec cette méthode. La différence la plus flagrante avec les résultats de la figure 4.5 est l'amélioration très nette du suivi de la commissure gauche pour $n=11$. L'amélioration du suivi de la commissure droite apparaît moins immédiatement. Cependant, les écarts par rapport à l'étiquetage manuel relevés dans le tableau 4.2 permettent de vérifier que le décalage des fenêtres conduit effectivement à un suivi plus précis des points P_1 et P_5. Les écarts correspondant aux autres points sont identiques à ceux du tableau 4.1 car leurs voisinages sont restés inchangés (voir figure 4.6). Comme nous l'avons fait remarquer plus haut, il peut arriver que les fenêtres des points P_3 et P_6 empiètent sur l'intérieur de la bouche. Il paraît donc logique d'effectuer également un décalage des voisinages associés à ces points. Cependant, nos tests ont montré que

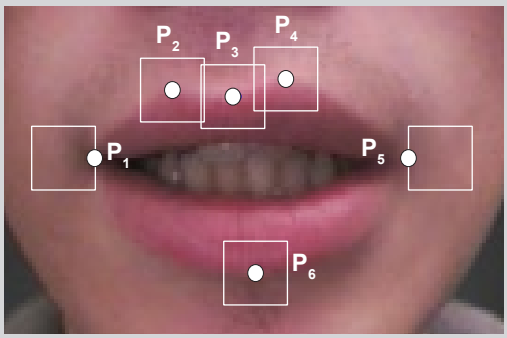


Fig. 4.6. Voisinages utilisés pour effectuer le suivi des points caractéristiques (repérés par des points). Les voisinages des commissures sont décalés de manière à ne pas empiéter sur l'intérieur de la bouche.

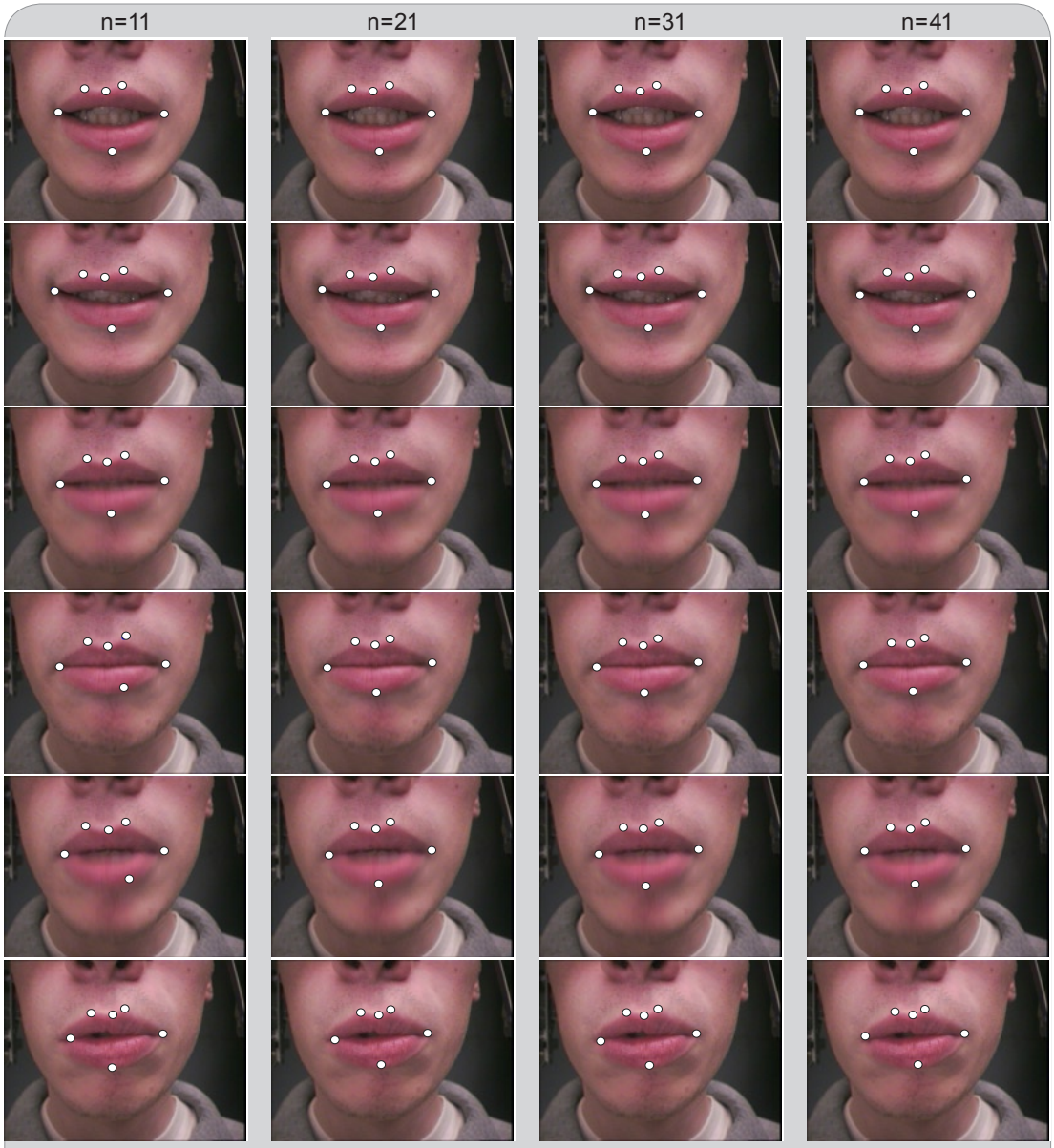


Fig. 4.7. Résultats du suivi des 6 points caractéristiques sur la séquence *Nico* pour différente tailles de fenêtre et en utilisant des fenêtres décalées pour les commissures.

n	P_1	P_2	P_3	P_4	P_5	P_6
11	8.8%	4.3%	4.3%	3.5%	4.7%	7.3%
21	7.2%	3.9%	4.3%	3.1%	4.7%	7%
31	6.8%	3.9%	4.4%	3.1%	4.7%	6.3%
41	6.7%	3.8%	4.6%	3.5%	4.6%	6.9%

Tab. 4.2. Ecarts moyens des points suivis sur la séquence *Nico* obtenus avec un décalage des fenêtres associées aux commissures. Les écarts sont donnés en pourcentage de la largeur de la bouche.

cela dégrade la qualité du suivi. En effet, dans ce cas, les fenêtres décalées ne contiennent plus de contours très marqués, ce qui rend le calcul des déplacements moins précis.

Finalement, n doit être suffisamment petit pour que le calcul soit rapide et pour éviter les phénomènes d'empiètement des points P_3 et P_6. De plus, il doit être suffisamment grand pour assurer une bonne précision aux suivis. **Nous avons choisi n=21.** Cette valeur permet d'effectuer des suivis en un temps raisonnable (voir figure 4.4) avec une bonne précision. De plus, elle permet d'éviter le phénomène d'empiètement tant que les épaisseurs des lèvres haute et basse sont supérieures à 10 pixels, ce qui est le cas de la plupart des bouches présentes dans notre base d'images.

4.2.3 Nécessité d'un recalage

A première vue, l'algorithme de Lucas-Kanade semble fournir des résultats corrects d'une image à la suivante. En réalité, les estimations de positions ne sont pas parfaites et l'accumulation progressive des erreurs mène souvent, après quelques images, à des résultats très imprécis. La figure 4.8 présente les suivis obtenus sur la séquence *Niko2*. Peu à peu, la précision des estimations se dégrade. On peut d'ailleurs noter que cette dégradation concerne surtout le point bas (à cause du problème d'ouverture mentionné plus haut) et les commissures. Malgré le décalage des fenêtres, elles ont tendance à «dériver» vers l'intérieur de la bouche. Bien que moins marqué, ce phénomène est également visible sur la figure 4.7.

Pour estimer la qualité du suivi, une solution relativement intuitive consiste à examiner la fonction coût ε donnée dans l'équation 4.2. Pour un suivi parfait, les contenus des fenêtres d'observation dans les 2 images adjacentes sont identiques et ε=0. A l'opposé, si ε est important, alors le suivi est mauvais. Cependant, en pratique il est très difficile de fixer un seuil sur ε permettant de valider ou non une estimation. Un point qui semble correctement suivi est parfois associé à une forte valeur de ε à cause d'un changement brusque de luminance ou à cause de

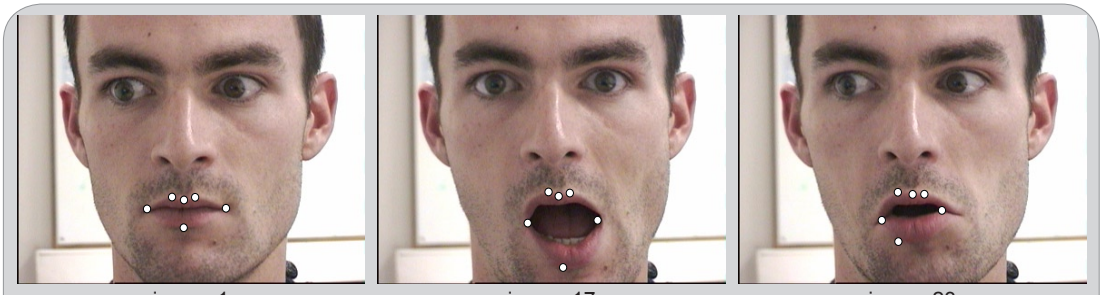


Fig. 4.8. Résultats de suivi sur la séquence *Niko2* en utilisant des fenêtres décalées pour les commissures. Après quelques images, les positions estimées deviennent peu fiables

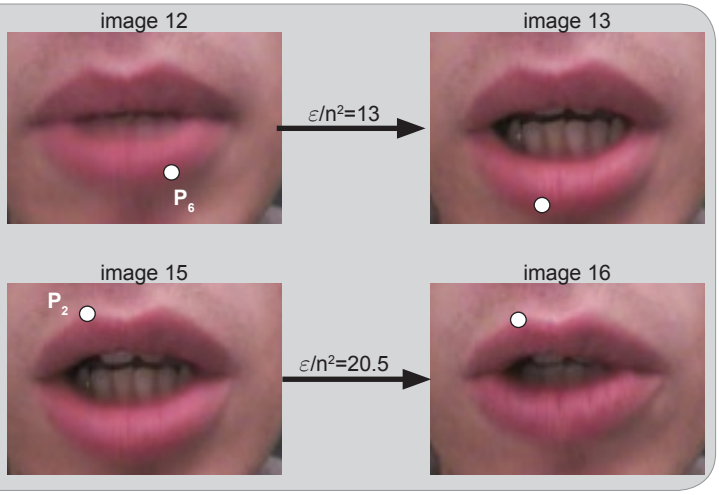


Fig. 4.9. La valeur de ε ne permet pas toujours d'estimer la qualité d'un suivi. Le suivi du haut est associé à une valeur de ε inférieure à celui du bas, alors qu'il est visiblement beaucoup moins précis.

l'apparition de rides. Par exemple, sur la figure 4.9, la valeur de ε pour l'estimation de mouvement de P_6 est inférieure à celle de P_2, alors que le suivi est manifestement moins bon.

Dans ces conditions, on peut donc considérer que l'algorithme de Lucas-Kanade est «aveugle», dans le sens où il est difficile de valider une estimation, et que ses prédictions doivent être affinées. La partie suivante présente un algorithme de recalage des points permettant de compenser les erreurs de suivi.

4.3 Recalage des points caractéristiques

Comme nous l'avons montré dans la partie précédente, les estimations fournies par l'algorithme de Lucas-Kanade doivent être affinées. Dans la suite de ce chapitre, ces estimations seront notées $\{P'_1(t),...,P'_6(t)\}$ pour l'image t. Les points recalés seront notés $\{P_1(t),...,P_6(t)\}$.

4.3.1 Recalage des points hauts et bas

4.3.1.1 Principe

Pour ajuster la position des points hauts et bas, nous détectons les contours supérieur et inférieur des lèvres en utilisant 2 *snakes* ouverts. Comme le montre la figure 4.10, les points les plus hauts du snake supérieur dans les voisinages de $P'_2(t)$ et $P'_4(t)$ sont les positions finales $P_2(t)$ et $P_4(t)$. De même, le point le plus bas du snake inférieur correspond au point bas recalé $P_6(t)$.

Comme nous l'avons déjà expliqué dans le chapitre 2 (partie 2.3), les *contours actifs* (ou *snakes*) sont des courbes $\boldsymbol{\nu}$ définies paramétriquement ($\boldsymbol{\nu}(s)=(x(s),y(s))$, où s est l'abscisse curviligne) qui peuvent se déformer progressivement de manière à s'approcher au plus près des contours d'un objet. Cette déformation est guidée par la minimisation d'une fonctionnelle

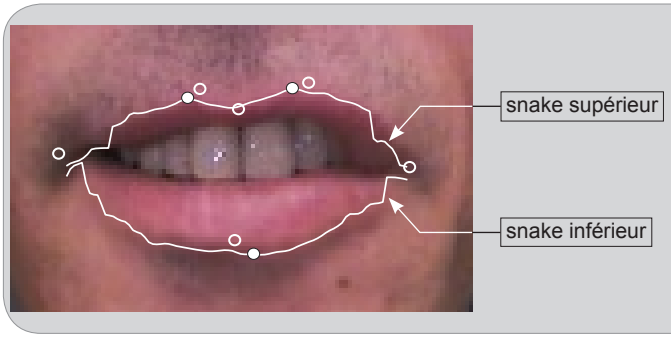


Fig. 4.10. Les points fournis par l'algorithme de Lucas-Kanade (cercles blancs) sont recalés en utilisant des snakes. Les points les plus hauts et les plus bas sur les snakes supérieur et inférieur sont les points recalés hauts et bas respectivement (points blancs).

d'énergie comprenant une énergie interne (permettant de régulariser le contour) et une énergie externe (attirant le *snake* vers les contours). La résolution des équations régissant le comportement du snake conduit à la relation dynamique suivante (se reporter à la partie 2.3.1 pour le détail des calculs) :

$$\boldsymbol{\nu}_i = (A+\gamma I_d)^{-1}(\gamma \boldsymbol{\nu}_{i-1} + \mathbf{F}_{\mathbf{ext}}(\boldsymbol{\nu}_{i-1})) \quad \text{(eq. 4.13)}$$

où $\boldsymbol{\nu}_i$ et $\boldsymbol{\nu}_{i-1}$ sont les positions du snake aux itérations i et $i-1$ respectivement. Le coefficient γ est appelé coefficient d'amortissement et contrôle la vitesse de déplacement du snake. $\boldsymbol{F}_{ext}$ est la force extérieure s'appliquant sur le contour actif. Dans le cas du recalage des points P_2 et P_4, le *snake* doit être attiré par le contour supérieur des lèvres. La force extérieure qui s'applique sur le snake du haut dérive donc du gradient hybride $\boldsymbol{R}_{top}$:

$$\boldsymbol{F}_{ext,\,haute} = \boldsymbol{\nabla}(|R_{top}|^2) \quad \text{(eq. 4.14)}$$

Pour le recalage du point bas P_6, *le snake* doit être attiré par le contour inférieur. La force extérieure basse dérive donc du gradient de la pseudo-teinte :

$$\boldsymbol{F}_{ext,\,basse} = \nabla(|\nabla[h]|^2) \quad \text{(eq. 4.15)}$$

A est la matrice de rigidité. Elle est de taille $N_s \times N_s$, où N_s est le nombre de points du snake, et est fonction des coefficients d'élasticité et de courbure α et β. Or, comme nous l'avons expliqué dans la partie 2.3, un des points délicats des *contours actifs* est le réglage de ces coefficients. Pour simplifier le problème, nous utilisons un snake sans force intérieure. Cela conduit à une relation dynamique plus simple ne comportant pas de matrice de rigidité :

$$\boldsymbol{\nu}_i = \boldsymbol{\nu}_{i-1} + \frac{1}{\gamma}\boldsymbol{F}_{ext}(\boldsymbol{\nu}_{i-1}) \quad \text{(eq. 4.16)}$$

Comme il n'existe plus de contrainte d'élasticité ni de courbure, les *snakes* obtenus sont bruités et irréguliers. Sur la figure 4.10, on peut remarquer que ce comportement est surtout marqué dans les zones de faible gradient. Mais notre but n'est pas d'extraire les contours en entier. Nous avons seulement besoin de les estimer dans les voisinages de $P'_2(t)$, $P'_4(t)$ et $P'_6(t)$. Dans ces zones, les gradients sont forts et les forces extérieures sont suffisamment importantes pour «plaquer» avec précision les *snakes* sur les contours.

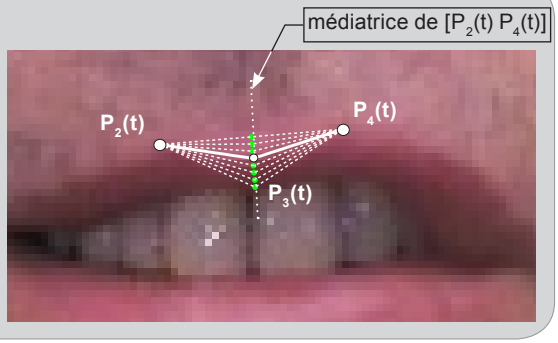


Fig. 4.11. Le point central haut $P_3(t)$ est trouvé en testant quelques candidats le long de la médiatrice de $[P_2(t)\ P_4(t)]$. Le meilleur point maximise le flux moyen de R_{top} à travers la ligne brisée $[P_2(t)\ P_3(t)\ P_4(t)]$.

Le point central haut $P_3(t)$ est trouvé en supposant qu'il est situé à égale distance des points $P_2(t)$ et $P_4(t)$. Comme le montre la figure 4.11, on teste donc une dizaine de positions dans le voisinage de $P'_3(t)$, le long de la médiatrice de $[P_2(t)\ P_4(t)]$. Le meilleur candidat maximise le flux moyen de $\boldsymbol{R}_{top}$ à travers la ligne brisée $[P_2(t)\ P_3(t)\ P_4(t)]$.

4.3.1.2 Initialisation

Pour déterminer la position initiale des *snakes*, on utilise la segmentation réalisée dans l'image précédente. Le modèle de lèvres obtenu à l'image *t-1* est étiré ou contracté de manière à coïncider avec les points caractéristiques estimés $P'_k(t)$. Pour cela, on applique une déformation linéaire aux cubiques de l'image *t-1*, notées $\gamma_k(t-1)$. Le déplacement de chacun de leur point est obtenu par une moyenne pondérée des déplacements des 2 points caractéristiques associés. Par exemple, les déplacements de P_1 et P_2 sont utilisés pour calculer le vecteur déplacement de chaque point Q de $\gamma_1(t-1)$:

$$\boldsymbol{d}_Q = \boldsymbol{d}_{P_1}\left(1-\frac{|P_1(i-1)Q(i-1)|}{|P_1(i-1)P_2(i-1)|}\right)+\boldsymbol{d}_{P_2}\left(1-\frac{|P_2(i-1)Q(i-1)|}{|P_1(i-1)P_2(i-1)|}\right) \quad \text{(eq. 4.17)}$$

où $\boldsymbol{d}_Q$, $\boldsymbol{d}_{P_1}$ et $\boldsymbol{d}_{P_2}$ sont les vecteurs déplacement des points Q, P_1 et P_2 respectivement :

$$\begin{cases} \boldsymbol{d}_Q = \overrightarrow{Q(i-1)Q(i)} \\ \boldsymbol{d}_{P_1} = \overrightarrow{P_1(i-1)P_1'(i)} \\ \boldsymbol{d}_{P_2} = \overrightarrow{P_2(i-1)P_2'(i)} \end{cases}$$

Cette transformation, illustrée par la figure 4.12, permet d'obtenir la courbe déformée $\gamma_1'(t)$. Elle est appliquée aux 4 courbes cubiques du modèle (les courbes résultantes sont notées $\gamma_k'(t)$) ainsi qu'à la ligne brisée $[P_2(t-1)\ P_3(t-1)\ P_4(t-1)]$. Sur la figure 4.13, on peut constater que le modèle déformé obtenu est relativement proche des contours des lèvres et peut être utilisé pour initialiser les 2 *contours actifs* nécessaires au recalage des points hauts et bas. La position initiale du snake supérieur est obtenue par un échantillonnage de $\gamma_1'(t)$, $\gamma_2'(t)$ et de la ligne brisée $[P'_2(t)\ P'_3(t)\ P'_4(t)]$, et celle du snake inférieur par l'échantillonnage de $\gamma_3'(t)$ et $\gamma_4'(t)$.

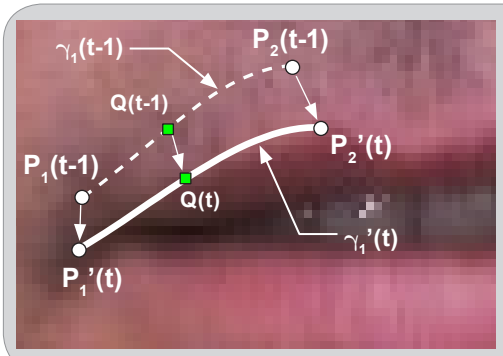


Fig. 4.12. La courbe γ_1(t-1) (en pointillé) est déformée de manière à passer par les points caractéristiques issus de l'algorithme de suivi (P_1'(t) et P_2'(t)). Pour cela, le déplacement de chaque point Q (■) est obtenu par une moyenne pondérée des déplacements de P_1 et P_2.

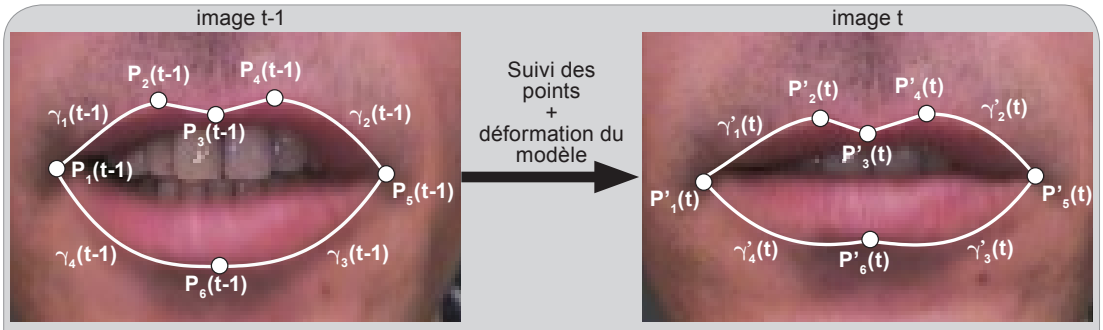


Fig. 4.13. La déformation des courbes cubiques γ_k(t-1) calculées à l'image *t-1* permet d'obtenir une première estimation du modèle de lèvre pour l'image *t*. Cette estimation est utilisée pour initialiser les snakes utilisés dans le recalage des points caractéristiques.

4.3.1.3 Mise en œuvre pratique

A partir de la position initiale obtenue par déformation du modèle, les *snakes* évoluent selon l'équation 4.16. Cependant, comme nous n'utilisons aucune contrainte d'élasticité ou de courbure, ils ont tendance à s'enrouler sur eux-mêmes. De même, leurs points se regroupent souvent dans les zones de fort gradient. La figure 4.14-a montre la position initiale des points du *snake* supérieur. Ces derniers sont régulièrement répartis (selon l'axe horizontal) le long des courbes du modèle déformé. Sur la figure 4.14-b, on peut observer qu'après quelques itérations les points sont absents de certaines zones du contour, ce qui nuit à la précision du processus de recalage des points caractéristiques. Cet artefact peut être supprimé en autorisant uniquement les mouvements verticaux. Dans ce cas, seule la composante verticale de la force extérieure est utilisée. La figure 4.14-c montre que cette précaution permet de conserver une répartition régulière des points. Une autre solution (utilisée notamment par Delmas dans [Delmas, 2000]) pour conserver une répartition régulière est le ré-échantillonage des points après chaque itération. Dans notre cas, cette technique ralentit l'algorithme et n'apporte pas de gain significatif sur la précision du recalage.

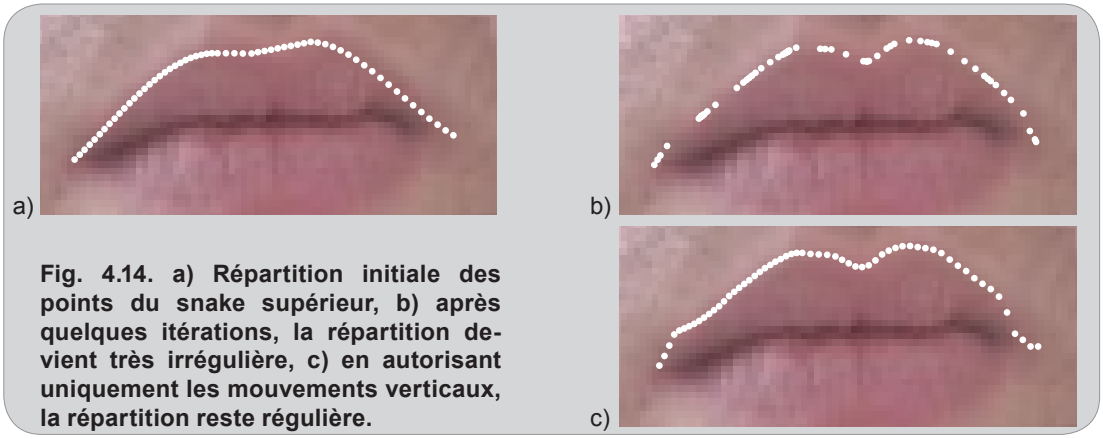


Fig. 4.14. a) Répartition initiale des points du snake supérieur, b) après quelques itérations, la répartition devient très irrégulière, c) en autorisant uniquement les mouvements verticaux, la répartition reste régulière.

Les *snakes* se déforment progressivement pour épouser les contours des lèvres. Après quelques itérations du processus de convergence, les positions des points caractéristiques sont ré-évaluées. Comme le montre la figure 4.10, les positions ajustées des points de l'*arc de Cupidon* (P_2 et P_4) sont obtenues en repérant les points les plus hauts des parties gauche et droite du snake supérieur. De même, la position ajustée du point bas (P_6) correspond au point le plus bas du snake inférieur. Si les positions ajustées sont proches des positions précédentes, le processus de recalage s'arrête. Sinon, la convergence des snakes continue pendant encore quelques itérations, puis les positions sont à nouveau ré-évaluées. En pratique, la position des points caractéristiques est ajustée tout les 10 itérations. Lorsque les distances entre les positions successives deviennent inférieures à 0.2 pixels, on considère que le recalage est effectué. On obtient ensuite le point central haut recalé $P_3(t)$ par la technique décrite à la fin de la partie 4.3.1.1 et illustrée par la figure 4.11.

Cette technique de recalage suppose que l'*arc de Cupidon* est bien dessiné. En effet, si la bouche est très arrondie, comme lors de la production du son **/u/**, les points $P_2(t)$ et $P_4(t)$ ont tendance à être très proches, voire à être confondus (voir figure 4.15). De même, si la bouche est étirée, la position horizontale de ces points est très incertaine. Théoriquement, l'algorithme

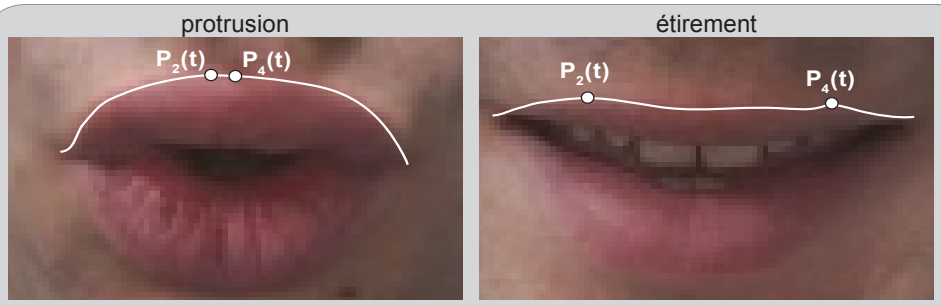


Fig. 4.15. Lorsque l'arc de Cupidon est mal dessiné, la position ajustée des points P_2 et P_4 (○) est incertaine. La ligne blanche symbolise la position finale du snake supérieur.

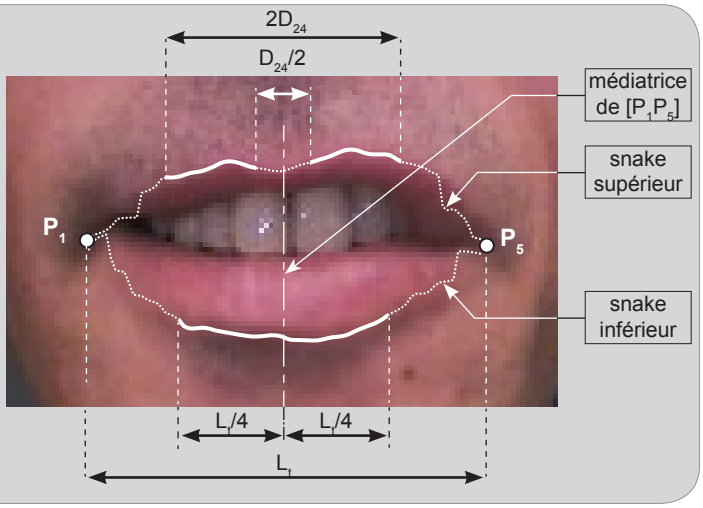


Fig. 4.16. De manière à éviter la dérive des points caractéristiques vers des positions aberrantes, seules les portions des snakes situées dans les zones autorisées (en trait épais) sont considérées lors du recalage.

de recalage est capable de retrouver les positions correctes dès que l'*arc de Cupidon* redevient visible. Cependant, lors de la segmentation des lèvres sur une séquence vidéo complète, les incertitudes passagères sur la position des points donnent l'impression que le contour supérieur «vibre» latéralement.

De manière à éviter cette dérive des points vers des positions aberrantes, il peut être intéressant de prendre en compte quelques contraintes morphologiques. Tout d'abord, la largeur de l'arc de Cupidon est à peu près constante. De plus, il est généralement situé sur l'axe de symétrie de la bouche. Nous avons donc inclus ces 2 hypothèses dans l'algorithme de recalage en définissant des «zones autorisées» sur les *snakes*, comme le montre la figure 4.16. Lors du recalage, seules les portions de *snake* présentes dans ces zones sont considérées. La largeur et la position des zones autorisées supérieures sont calculées en utilisant la position des commissures et la distance entre $P_2(0)$ et $P_4(0)$ dans la première image de la séquence (cette distance est notée D_{24}). La zone autorisée inférieure est définie en utilisant uniquement la position courante des commissures. On note L_t la largeur de la bouche à l'instant t.

4.3.2 Recalage des commissures

La position des commissures ne peut pas être ajustée en utilisant les *contours actifs* car elles sont généralement situées dans des zones de faible gradient. La technique que nous employons pour les recaler est donc assez proche de celle que nous avons utilisé pour déterminer leur position initiale dans la première image, dans le sens où elles sont trouvées en utilisant la continuité des contours (voir partie 3.4.2).

Nous considérons quelques commissures possibles le long de la ligne des minima de luminance L_{mini}. Comme nous disposons d'estimations approximatives des commissures ($P'_1(t)$ et $P'_5(t)$), seul un nombre réduit de candidats est utilisé. Typiquement, un bon recalage peut être obtenu en considérant les 4 ou 5 voisins les plus proches de $P'_1(t)$ et $P'_5(t)$ sur L_{mini}. Il s'agit ensui-

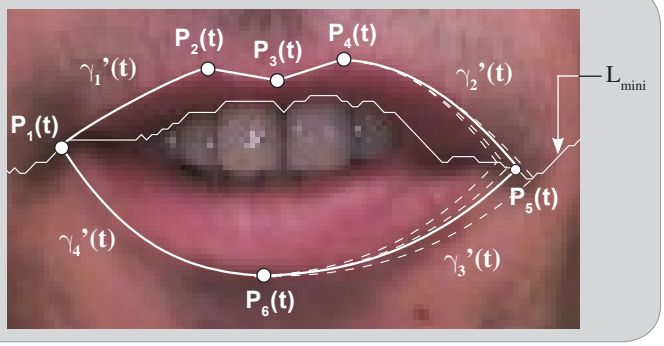


Fig. 4.17. Pour déterminer la commissure optimale, 4 positions possibles sont testées le long de L_{mini}. Les lignes en pointillés sont les courbes correspondantes obtenues par la déformation des cubiques de l'image précédente $\gamma_k(t\text{-}1)$.

te d'associer à chacun de ces candidats un couple de courbes approchant au mieux le contour des lèvres. Lors de la détection des commissures dans l'image initiale, ces courbes étaient obtenues en utilisant quelques points supplémentaires placés le long des contours (voir partie 3.4.2). Ici, nous disposons de la segmentation effectuée dans l'image précédente. Or, la forme de la bouche change peu d'une image à l'autre. La déformation des cubiques $\gamma_k(t\text{-}1)$ de l'image précédente par l'équation 4.17 fournit donc de bonnes approximations du contour, comme le montre la figure 4.17. Ainsi, à chaque commissure testée est associé un couple de courbes $\gamma_k'(t)$ obtenues par déformation des cubiques de l'image précédente. Comme dans la partie 3.4.2, les commissures optimales $P_1(t)$ et $P_5(t)$ maximisent ϕ_{total} sur les côtés gauche et droit respectivement (voir équation 3.14). La figure 4.17 illustre le processus de recalage de la commissure droite.

4.4 Détermination du contour final

4.4.1 Calcul des cubiques

En général, les courbes déformées $\gamma_k'(t)$ obtenues dans la partie précédente suivent assez bien le contour des lèvres, comme le montre la figure 4.17. Cependant, ces courbes ne sont pas des cubiques. Par conséquent, de manière à conserver la structure et la cohérence du modèle, la dernière étape de la segmentation est le calcul des cubiques associées aux points caractéristiques recalés $P_k(t)$. Ces cubiques sont notées $\gamma^c_k(t)$

Tout comme pour la détermination du modèle dans la première image de la séquence (voir chapitre précédent), les meilleures cubiques sont calculées en examinant des flux moyens de vecteurs gradient les traversant. Le contour supérieur de la bouche est obtenu par la maximisation des flux moyens $\phi_{top,1}$ et $\phi_{top,2}$ (voir équation 3.13). De même, le contour inférieur est obtenu par la minimisation des flux moyens $\phi_{low,3}$ et $\phi_{low,4}$. Chacune des cubiques est décrite de manière unique si ses 4 paramètres sont connus. Ici, nous connaissons déjà la position des points caractéristiques par lesquels ces dernières doivent passer. De plus, nous savons également que leurs dérivées s'annulent aux points $P_2(t)$, $P_4(t)$ et $P_6(t)$. Ces contraintes fournissent 3 équations qui font passer le nombre de paramètres à estimer de 4 à 1 pour chaque cubique.

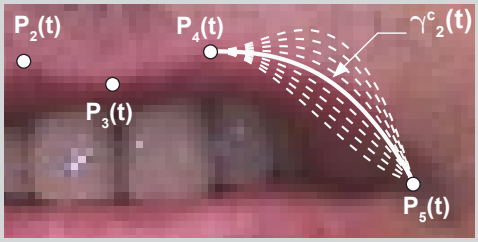


Fig. 4.18. Quelques cubiques avec différentes pentes aux commissures sont testées (ligne en pointillés). La meilleure (en trait plein) est $\gamma^c_2(t)$. Elle maximise $\phi_{top,2}$.

Le dernier paramètre peut être déterminé si les pentes sur les commissures sont connues. On teste donc différentes valeurs de pente. La cubique associée à la pente optimale maximise le flux moyen. La figure 4.18 illustre ce processus pour l'extraction de $\gamma^c_2(t)$, la cubique décrivant le côté supérieur droit. Les cubiques testées (en pointillés) passent par $P_4(t)$ et $P_5(t)$, et ont des pentes différentes en $P_5(t)$. La meilleure (en trait plein) maximise le flux moyen $\phi_{top,2}$.

De manière à diminuer le nombre de pentes testées, nous utilisons les résultats obtenus dans l'image précédente. Comme nous supposons que la bouche se déforme lentement, les pentes entre 2 images consécutives sont proches. Par conséquent, on considère uniquement des valeurs de pentes proches de celles de l'image précédente. En pratique, si l'on note $\rho_k(t\text{-}1)$ la pente angulaire (exprimée en radians) de la cubique $\gamma_k(t\text{-}1)$ dans l'image *t-1*, nos essais ont montré qu'on obtient de bons résultats en testant une dizaine de pentes réparties régulièrement dans l'intervale $[\rho_k(t\text{-}1)\text{-}\pi/6 ; \rho_k(t\text{-}1)+\pi/6]$. Au-delà de l'accroissement de la vitesse de l'algorithme, cette utilisation des résultats obtenus à l'image précédente permet d'améliorer la robustesse. En effet, dans le cas où le contour des lèvres devient peu visible, une recherche exhaustive des pentes dans l'intervale $[-\pi/2 ; \pi/2]$ conduit souvent à des résulats aberrants (voir figure 4.19). A l'opposé, l'utilisation d'un intervalle restreint centré sur la pente obtenue à l'image précédente permet de conserver une forme cohérente qui se déforme de manière continue dans le temps.

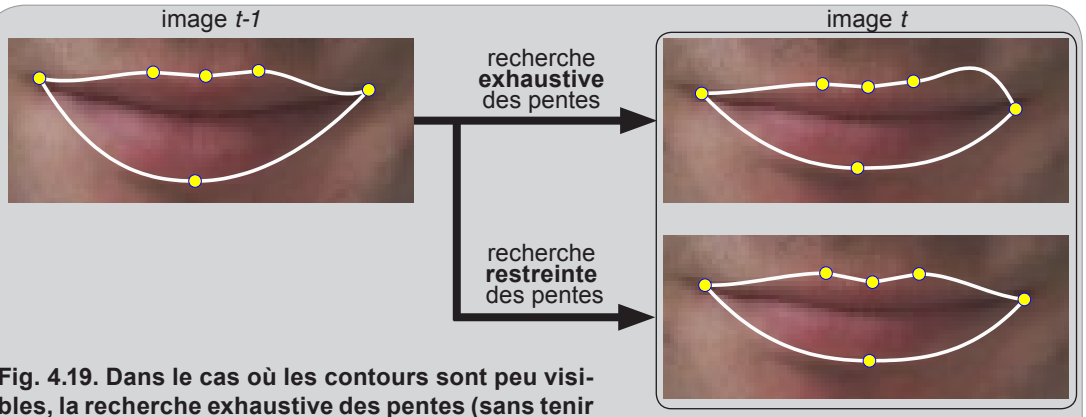


Fig. 4.19. Dans le cas où les contours sont peu visibles, la recherche exhaustive des pentes (sans tenir compte des pentes obtenues à l'image précédente) peut conduire à des résultats aberrants. A l'opposé, une recherche menée sur un intervalle restreint permet de conserver une forme cohérente.

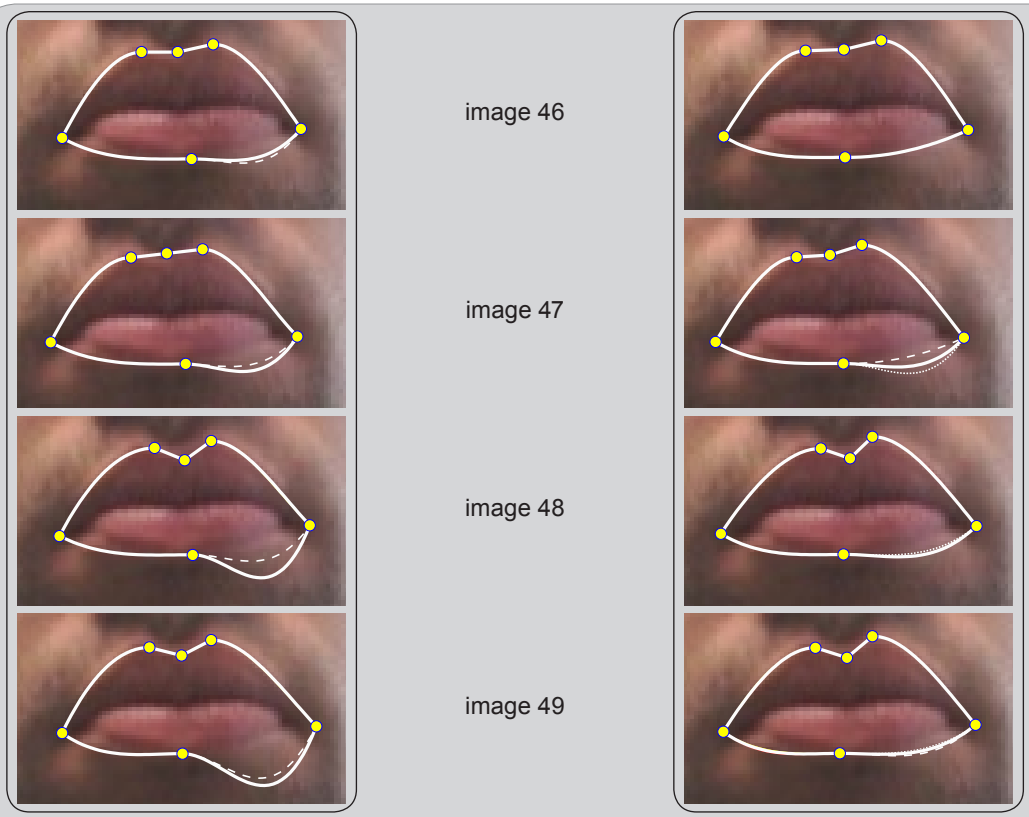


Fig. 4.20. L'utilisation d'une moyenne pondérée pour calculer les courbes du modèle permet de stabiliser le contour. Les courbes déformées sont représentées en pointillés longs. a) segmentation obtenue en utilisant uniquement des cubiques, b) en utilisant une moyenne des cubiques et des courbes déformées (les cubiques sont représentées par des pointilés courts).

4.4.2 Stabilisation du contour

A priori, la segmentation des lèvres dans l'image t est terminée lorsque les cubiques $\gamma^c_k(t)$ constituant le modèle ont été déterminées. Si l'on pose $\gamma_k(t) = \gamma^c_k(t)$, les résultats obtenus sur des séquences vidéo sont en général de bonne qualité. Cependant, dans le cas de contours peu marqués ou sous-éclairés, il arrive que les courbes du modèle «vibrent» ou prennent temporairement des formes aberrantes, malgré la restriction des intervalles de recherche des pentes expliquée précédemment. Par exemple, sur la figure 4.20-a, la cubique correspondant au contour inférieur droit s'éloigne peu à peu des lèvres pour aller se positionner vers la limite d'une zone d'ombre.

Pour atténuer ce phénomène, nous considérons que les courbes finales sont les moyennes des cubiques $\gamma^c_k(t)$ et des courbes déformées $\gamma'_k(t)$. Cependant, si la cubique est plus proche

du contour que la courbe déformée, alors elle doit avoir un poid plus important. La moyenne est donc pondérée par les valeurs des flux moyens :

$$\gamma_k(t) = \frac{\gamma_k^c(t)\,\phi(\gamma_k^c(t)) \;+\; \gamma'_k(t)\,\phi(\gamma'_k(t))}{\phi(\gamma_k^c(t)) \;+\; \phi(\gamma'_k(t)} \qquad \text{(eq. 4.18)}$$

où $\phi(\gamma^c_k(t))$ et $\phi(\gamma'_k(t))$ sont les flux moyens à travers $\gamma^c_k(t)$ et $\gamma'_k(t)$. On peut observer sur la figure 4.20-b, que l'utilisation de cette moyenne pondérée permet d'éviter la *dérive* du contour inférieur droit obtenue sur l'image 4.20-a. D'une manière générale, ce calcul des $\gamma_k(t)$ conduit à des contours plus stables se déformant de manière continue. De plus, même si les courbes $\gamma_k(t)$ ne sont pas des cubiques, elles s'en approchent suffisamment pour conserver la structure du modèle. Dès lors, on parlera d'un modèle à structure «quasi-cubique».

4.5 Résultats

Dans cette partie, nous présentons et analysons les résultats obtenus avec notre algorithme complet. Le corpus que nous utilisons est le même que celui qui a permi de tester la méthode de segmentation statique du chapitre précédent. Il contient environ 2000 images réparties en 30 séquences de 22 locuteurs. Les conditions de prise de vue sont relativement hétérogènes, tant du point de vue du cadrage que de l'éclairement. Cependant, il n'y a aucun cas d'occultation et les deux commissures sont toujours visibles. De plus, dans la première image de chaque séquence, la bouche doit être «d'apparence moyenne» (voir chapitre précédent) pour que l'algorithme de segmentation statique puisse déterminer le contour initial.

4.5.1 Pertinence du modèle

La figure 4.21 présente quelques résultats représentatifs de notre algorithme. On peut remarquer que la grande flexibilité du modèle permet de reproduire avec précision un large panel de formes de bouche. Par exemple, dans la séquence *Vero*, les lèvres ont une forme relativement prototypique, l'arc de Cupidon est bien marqué et les contours sont réguliers. Le modèle s'adapte parfaitement au contour et reproduit même la très légère asymétrie de la lèvre inférieure dans la quatrième image. Dans la séquence *Benny*, les conditions de prise de vue sont identiques à celles de *vero*; mais les contours inférieurs sont un peu moins marqués et plus asymétriques. Ensuite, on peut observer que le modèle est capable de suivre le contour lors d'une protrusion (séquence *Seb2*) ou lors d'un étirement (séquence *Nat*). De plus, même dans le cas d'une ouverture extrême (séquence *Alek2*), la segmentation précise reste possible. Enfin, la très grande flexibilité du modèle permet également la segmentation lorsque la déformation de la bouche est volontairement exagérée, comme lors d'une grimace (séquence *Niko2*). Il est d'ailleurs intéressant de noter que, contrairement à la *segmentation statique*, le domaine d'application de l'algorithme de *segmentation dynamique* n'est pas limité aux lèvres «d'apparence

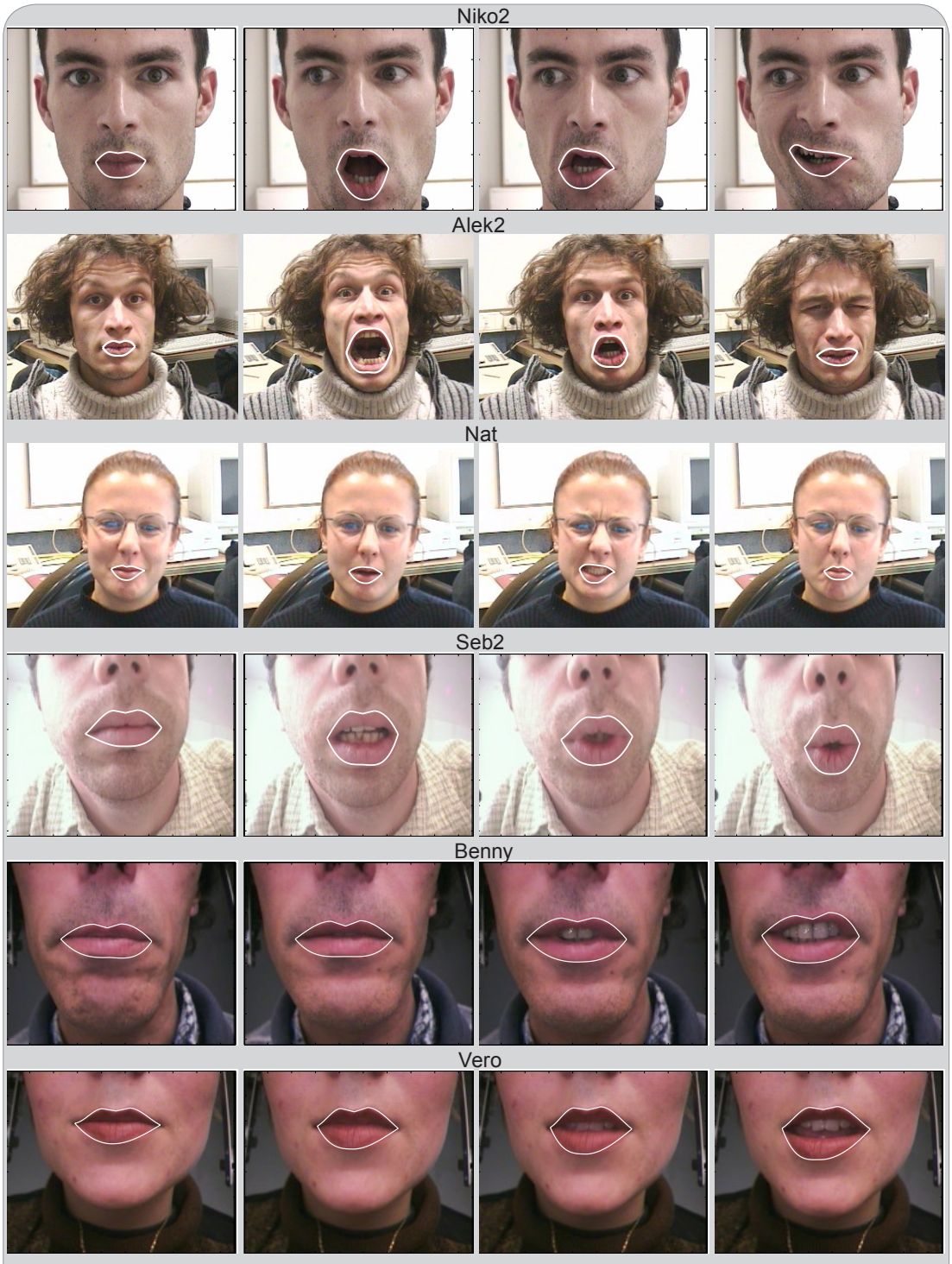


Fig. 4.21. Quelques résultats représentatifs de notre algorithme. Les segmentations de gauche sont les premières de chaque séquence et sont obtenues par l'algorithme de segmentation statique. Les autres sont obtenues par suivi.

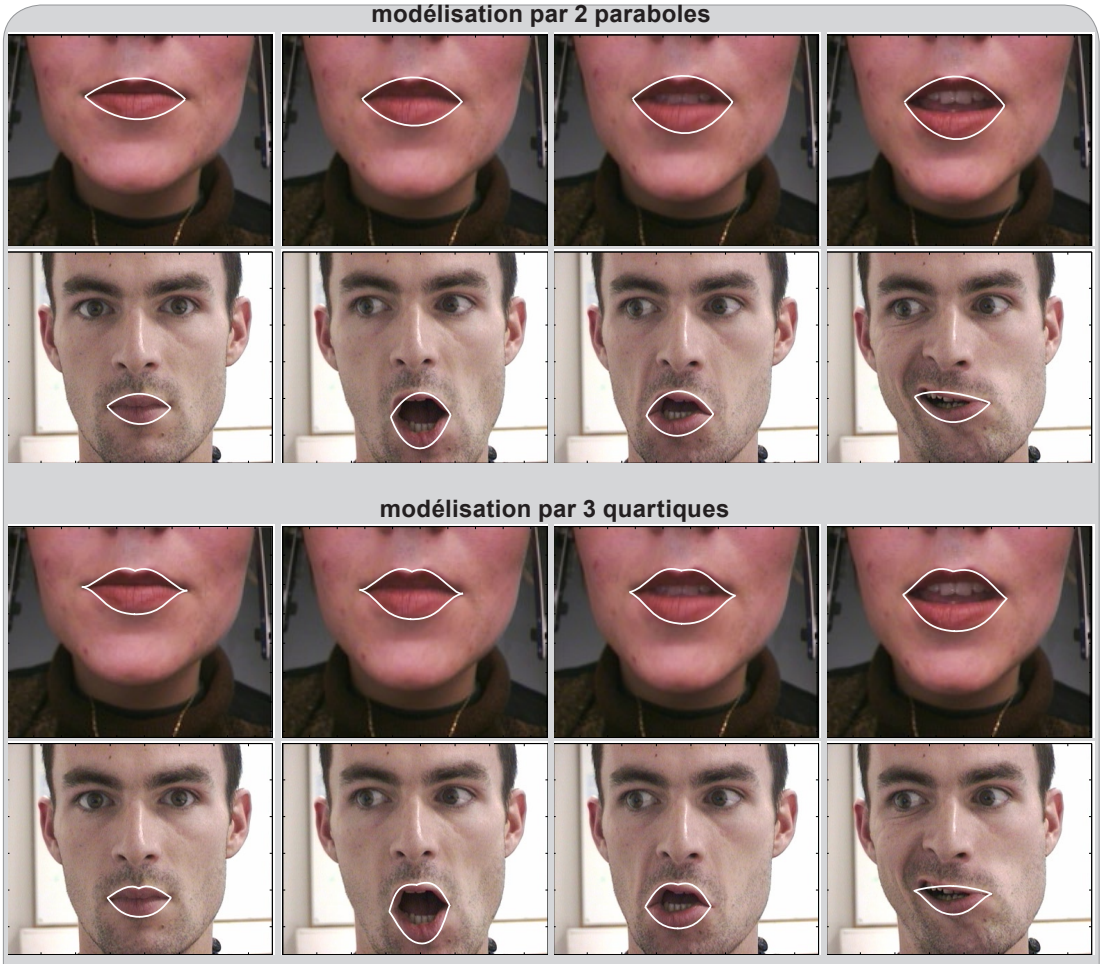


Fig. 4.22. Exemples de segmentations réalisées en utilisant un modèle bi-parabolique et un modèle constitué de quartiques.

moyenne» (voir chapitre précédent). Enfin, on peut constater que, même si ses courbes ne sont pas rigoureusement des cubiques, le modèle conserve sa structure tout au long de la séquence.

La figure 4.22 donne quelques résultats comparatifs de segmentations effectuées avec 2 modèles paramétriques classiques sur les séquences *Vero* et *Niko2*. Le premier modèle est constitué de 2 paraboles (voir figure 2.13-a) et est utilisé notamment dans [Botino, 2002], [Tian *et al.*, 2000] et [Rao and Mersereau, 1995]. Le deuxième modèle est un peu plus complexe et représente le contour par 3 quartiques [Yuille *et al.*, 1992][Hennecke *et al.*, 1994]. Il est présenté à la figure 2.12. Les méthodes pour faire converger ces modèles sont nombreuses. Mais comme notre but est d'illustrer le gain de précision apporté par l'utilisation de notre modèle cubique, les adaptations aux contours ont été faites manuellement. On peut tout d'abord constater que le modèle bi-parabolique ne permet pas de représenter l'arc de Cupidon, qui est pourtant très bien marqué sur la séquence *vero*. De plus, comme le fait remarquer Tian dans [Tian *et al.*,

2000], il est trop rigide pour s'adapter fidèlement aux bouches asymétriques. Dès lors, il ne donne qu'une description très sommaire de la forme des lèvres. L'utilisation d'un modèle à 3 quartiques améliore un peu la précision et le réalisme de la segmentation. Il permet de représenter approximativement l'arc de Cupidon. Cependant, tout comme le modèle bi-parabolique, il ne peut s'adapter correctement aux formes asymétriques. Finalement, le modèle cubique que nous proposons apporte une amélioration très nette de la précision par rapport aux modèles de lèvres proposés jusqu'ici.

4.5.2 Précision

Pour évaluer quantitativement la précision des contours obtenus, il faut disposer d'une «vérité de terrain», c'est-à-dire d'un ensemble d'images pour lesquelles les contours ont été repérés manuellement par un expert. De plus, le nombre d'images incluses dans cet ensemble de référence doit être suffisamment important pour que les mesures de précision soient significatives. Selon nous, la méthode de saisie manuelle la plus rapide est basée sur les *splines*. Comme le montre la figure 4.23, elle permet de tracer un contour relativement lisse et régulier à partir de quelques points disposés autour des lèvres. Nous avons estimé qu'un minimum de 18 points était nécessaire pour décrire précisément la forme des lèvres. Par conséquent, si l'on veut obtenir une base étiquetée suffisamment importante (plusieurs centaines d'images), il faut saisir un très grand nombre de points. L'ampleur de la tâche ainsi que le peu de temps dont nous disposions ne nous ont pas permis de constituer une telle base. Dans sa thèse, Daubias a proposé une méthode originale d'étiquetage basée sur la répétition d'une même phrase par un locuteur, avec et sans maquillage bleu sur les lèvres [Daubias, 2002]. L'utilisation des informations acoustiques permet ensuite d'aligner les deux séquence et d'extraire le contour sur les images *naturelles*. Cette technique semble très efficace car Daubias l'a utilisée pour étiqueter plus de 5000 images ! Malheureusement, nous n'avons pas eu le temps de l'expérimenter.

Afin de faciliter la création de la base étiquetée, nous avons uniquement évalué la précision du suivi de points réalisé par notre algorithme. Pour cela, nous avons repéré manuellement les 6 points caractéristiques sur environ 300 images provenant de 8 séquences. Pour chaque image i, le point de référence k est noté $P_{k,ref}(i)$. Le point correspondant detecté par notre algorithme est noté $P_{k,suivi}(i)$. L'erreur moyenne de suivi pour chacun des points caractéristiques est notée $\varepsilon_{k,suivi}$ et vaut :

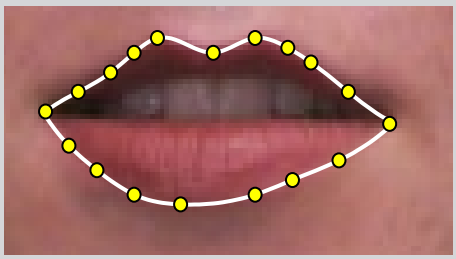


Fig. 4.23. Extraction d'un contour de référence par les *splines*.

Tab. 4.3. Erreurs moyennes (normalisées par rapport à la largeur de la bouche) sur la position des points caractéristiques pour l'algorithme de Lucas-kanade (1ère ligne), pour notre algorithme (2ème ligne) et pour un opérateur humain (3ème ligne).

	P_1	P_2	P_3	P_4	P_5	P_6
$\varepsilon_{k,kanade}$	9.5%	4.3%	4.5%	4.6%	7.4%	8.7%
$\varepsilon_{k,suivi}$	4.5%	2.9%	2.3%	2.4%	3.8%	4.1%
$\varepsilon_{k,humain}$	1.5%	1.1%	1.1%	1.2%	1.6%	1.6%

$$\varepsilon_{k,suivi} = \frac{1}{T}\sum_{i=1}^{T}\frac{|\overrightarrow{P_{k,ref}(i)P_{k,suivi}(i)}|}{|\overrightarrow{P_{1,ref}(i)P_{5,ref}(i)}|} \qquad \text{(eq. 4.19)}$$

où T est le nombre total d'images étiquetées. Les erreurs de suivi sont normalisées et s'expriment en pourcentage de la largeur de la bouche. De plus, nous mesurons également les erreurs $\varepsilon_{k,kanade}$ des estimations fournies par l'algorithme de Lucas-Kanade seul :

$$\varepsilon_{k,kanade} = \frac{1}{T}\sum_{i=1}^{T}\frac{|\overrightarrow{P_{k,ref}(i)P_{k,kanade}(i)}|}{|\overrightarrow{P_{1,ref}(i)P_{5,ref}(i)}|} \qquad \text{(eq. 4.20)}$$

où $P_{k,kanade}(i)$ est l'estimation de la position du point k dans l'image i, fournie par l'algorithme de Lucas-Kanade. Enfin, nous comparons $\varepsilon_{k,kanade}$ et $\varepsilon_{k,suivi}$ à l'erreur effectuée par l'expert lors de la saisie manuelle des points. Pour cela, les 6 points caractéritiques de 30 images selectionnées aléatoirement parmi les images de test sont saisis une dizaine de fois. Comme le marquage manuel n'est pas parfait, il existe une légère dispersion des points que nous mesurons en calculant $\varepsilon_{k,humain}$:

$$\varepsilon_{k,humain} = \frac{1}{T}\sum_{i=1}^{T}\left(\frac{1}{M}\sum_{m=1}^{M}\frac{|\overrightarrow{P_{k,ref}(i)P_{k,humain}(i,m)}|}{|\overrightarrow{P_{1,ref}(i)P_{5,ref}(i)}|}\right) \qquad \text{(eq. 4.21)}$$

où M est le nombre de sessions de saisie manuelles (M=10). $P_{k,humain}(i,m)$ est le point caractéristique k saisi lors de la session m sur l'image i. De plus, ici $P_{k,ref}(i)$ est la moyenne des saisies manuelles pour le point k sur l'image i :

$$P_{k,ref}(i) = \frac{1}{M}\sum_{m=1}^{M}P_{k,humain}(i,m) \qquad \text{(eq. 4.22)}$$

Le tableau 4.3 présente les valeurs des erreurs moyennes normalisées de notre algorithme, de l'algorithme de Lucas-Kanade et de la saisie manuelle. On peut constater que le recalage des points permet, dans une certaine mesure, de compenser les erreurs réalisées par l'algorithme de Lucas-Kanade. Cette amélioration est plus importante pour les commissures (P_1 et P_5) et pour le point bas (P_6). En effet, comme nous l'avons expliqué dans la partie 4.3.2, le suivi des commissures par l'algorithme de Lucas-Kanade est difficile car elles sont situées dans des zones fortement déformables. De même, le suivi du point bas est souvent imprécis à cause du *phénomène d'ouverture* (voir partie 4.2.2) et de la rapidité des mouvements de la lèvre inférieure.

Finalement, l'algorithme de recalage que nous proposons permet d'estimer la position des points caractéristiques avec une précision comparable à celle d'une saisie manuelle. Cela le

Fig. 4.24. Résultats de segmentation dans le cas d'une illumination variable, avec recalage des points (1ère ligne) et en utilisant seulement l'algorithme de Lucas-Kanade (2ème ligne).

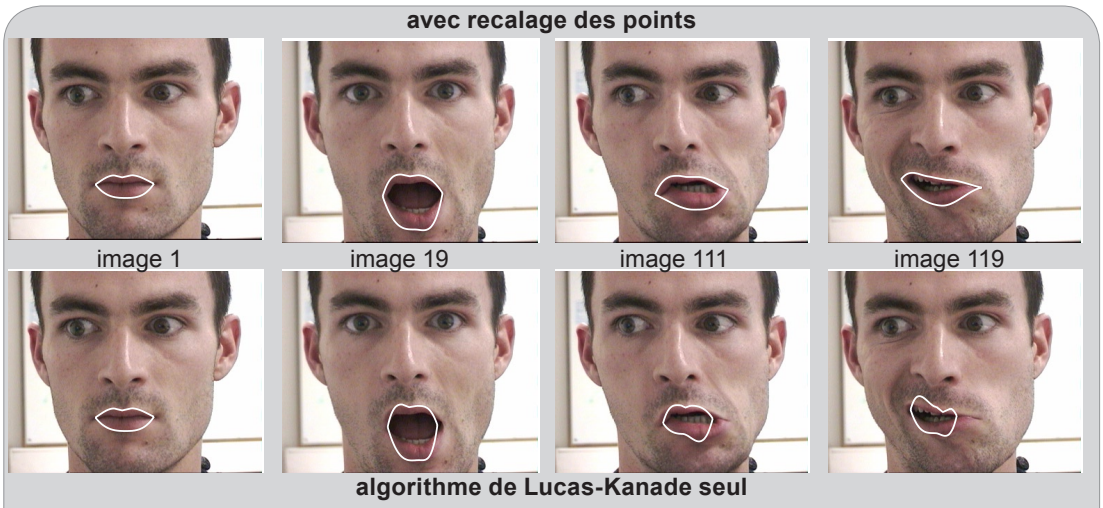


Fig. 4.25. Résultats de segmentation dans le cas d'une forte déformation, avec recalage des points (1ère ligne) et en utilisant seulement l'algorithme de Lucas-Kanade (2ème ligne).

rend utilisable dans des applications qui nécessitent une détection précise de ces points.

4.5.3 Robustesse

Les figures 4.24 et 4.25 illustrent le gain de robustesse apporté par notre algorithme par rapport à l'algorithme Lucas-Kanade. Elles présentent les segmentations effectuées sur des

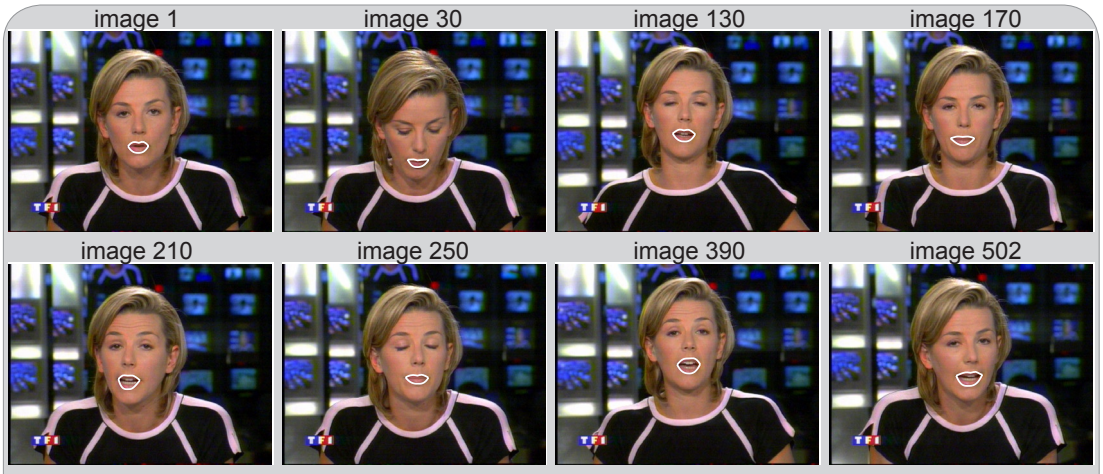


Fig. 4.26. Résultat de segmentation sur une longue séquence. Même après quelques centaines d'images, la segmentation reste correcte.

séquences particulièrement délicates. La première (figure 4.24) traite le cas d'une illumination très variable. Dans la deuxième (figure 4.25), la bouche du locuteur subit des déformations extrêmes. Pour chaque figure, les segmentations effectuées avec et sans recalage sont données sur la première et deuxième ligne respectivement. De plus, les contours ont tous été extraits en utilisant la méthode décrite dans la partie 4.4. On peut constater sur les deux séquences que les résultats fournis par l'algorithme de Lucas-Kanade deviennent rapidement aberrants. Comme nous l'avons dit dans la partie 4.2.3, cet algorithme est «aveugle», dans le sens où il est difficile d'évaluer l'exactitude de ses prédictions. Les erreurs faites sur la position des points s'accumulent et conduisent peu à peu à des résultats très approximatifs. Sur les figures 4.24 et 4.25, ce phénomène est particulièrement visible sur la commissure droite. A l'opposé, les segmentations effectuées avec recalage restent fiables, même dans les cas difficiles présentés. La compensation systématique des erreurs de suivi maintient le modèle près du contour des lèvres et permet d'envisager la segmentation de séquences très longues, comme le montre la figure 4.26. Enfin, au-delà des variations d'éclairement et des formes de bouches atypiques, notre algorithme est également robuste vis-à-vis des rotation planes (voir figure 4.27).

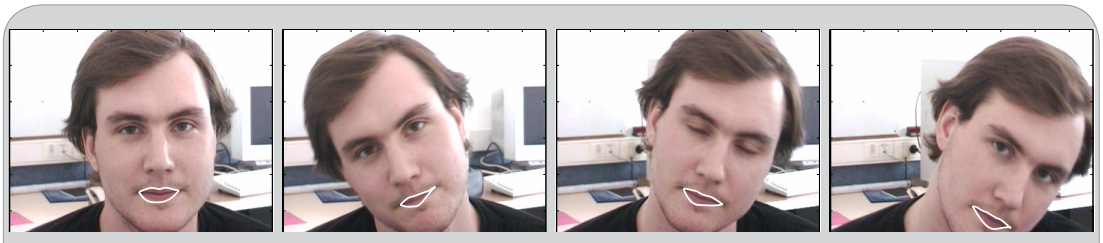


Fig. 4.27. Robustesse vis-à-vis de la rotation plane.

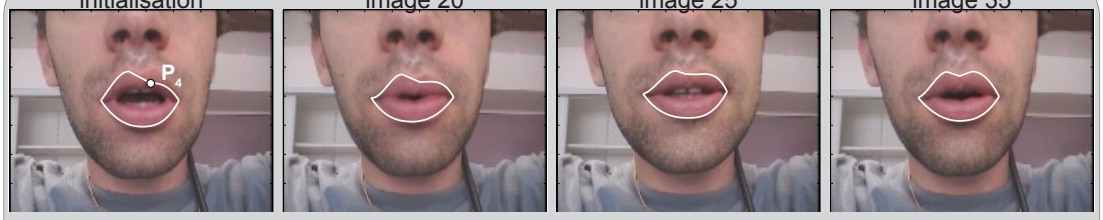


Fig. 4.28. Compensation d'une erreur d'initialisation.

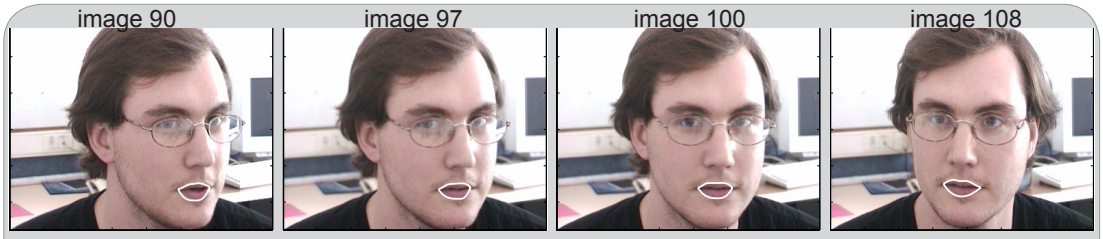


Fig. 4.29. Compensation d'une erreur due à un reflet passager sur la lèvre inférieure.

En plus d'assurer un suivi robuste des points, le recalage permet également de compenser des erreurs plus importantes apparaissant lors de l'initialisation. Par exemple, sur la figure 4.28, le point droit de l'*arc de Cupidon* (P_4) a été mal détecté à l'initialisation, ce qui conduit à une estimation fausse du contour supérieur droit. Cette erreur, déjà mentionnée dans la partie 3.5, est due au fait que la lèvre supérieure est trop *pentue* pour que le *jumping snake* puisse suivre son contour correctement. On peut constater sur les images suivantes que l'algorithme de recalage ramène progressivement P_4 vers une position correcte. De même, les erreurs apparaissant lors du suivi peuvent également être corrigées. Sur la figure 4.29, le reflet apparaissant sur le coté droit de la lèvre inférieure conduit à une segmentation imprécise pour l'image 90. Comme le reflet est toujours présent sur l'image 97, l'algorithme de recalage ne peut corriger la position de la commissure droite. Cependant, dès que le locuteur tourne à nouveau la tête vers la caméra, le reflet disparait et la commissure est ramenée vers une position correcte.

4.5.4 Vitesse

L'utilisation des informations temporelles apporte un gain de vitesse significatif par rapport à l'extraction directe mise en oeuvre dans le chapitre précédent. Les tableaux 4.4 et 4.5 présentent les temps de calcul moyens détaillés pour la *segmentation statique* et pour la *segmentation dynamique*, sur des images au format QCIF (de taille 144×192). Ces temps ont été obtenus sur un ensemble représentatif de notre corpus contenant des images pour lesquelles les bouches ont des largeurs comprises entre 55 et 95 pixels. L'algorithme a été implanté sous MATLAB et exécuté sur un PIV 2.4 GHz. Dans cette configuration, la cadence de traitement est proche de 4 images par seconde. On peut raisonnablement penser que l'utilisation d'une carte dédiée ou

calcul des gradients	0.04 s
convergence du *jumping snake*	0.29 s
détection des points hauts et bas	0.04 s
extraction du contour	0.35 s
Temps total pour l'initialisation	**0.72 s**

Tab. 4.4. Temps moyens d'exécution détaillés pour la segmentation statique.

calcul des gradients	0.04 s
suivi des points caractéristiques	0.08 s
recalage	0.07 s
détermination du contour	0.07 s
Temps total pour le suivi	**0.26 s**

Tab. 4.5. Temps moyens d'exécution détaillés pour la segmentation dynamique.

qu'une implantation en C++ pourraient permettre d'atteindre un traitement en temps réel.

4.5.5 Limites de la méthode

Dans le chapitre 3, nous avons défini les limites d'application de l'algorithme de segmentation statique. Nous avons déterminé que, dans la première image, les lèvres doivent être «d'apparence moyenne», c'est-à-dire au repos et sans expression forcée. De plus, leur contours doivent être bien dessinés et leur largeur doit être comprise entre 40 et 100 pixels. Les parties précédentes ont montré que la prise en compte des informations temporelles permet de traiter des formes beaucoup plus variées dans des conditions d'éclairement moins calibrées. Le locuteur doit donc se «tenir tranquille» et être bien éclairé lors de l'initialisation, et peut ensuite bouger plus librement une fois que le suivi a commencé. Cependant, dans certaines conditions particulières, l'algorithme de segmentation dynamique donne des résultats partiellement, ou totalement, faux.

Tout d'abord, il existe peu de **restrictions géométriques**. Comme nous l'avons déjà montré, le modèle est très flexible et peut s'adapter à des formes très variées. La seule limite que nous avons observée concerne la pente des courbes sur les commissures. Comme ces courbes sont quasiment des cubiques, une forte pente se traduit par des dépassements verticaux très importants. Par exemple, sur la figure 4.30, le sourire forcé du locuteur impose une pente infinie

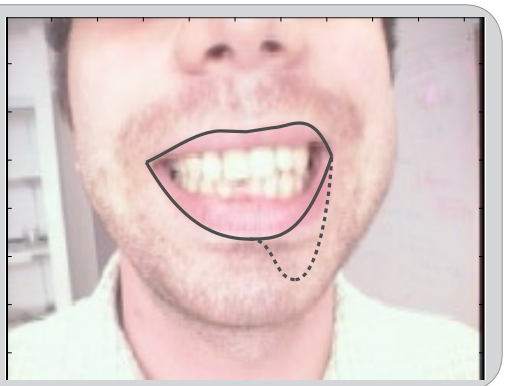


Fig. 4.30. Lorsque la pente des lèvres sur les commissures est trop importante, les cubiques ne peuvent suivre correctement le contour. En trait plein : la courbe choisie. En pointillés : une cubique de forte pente sur la commissure

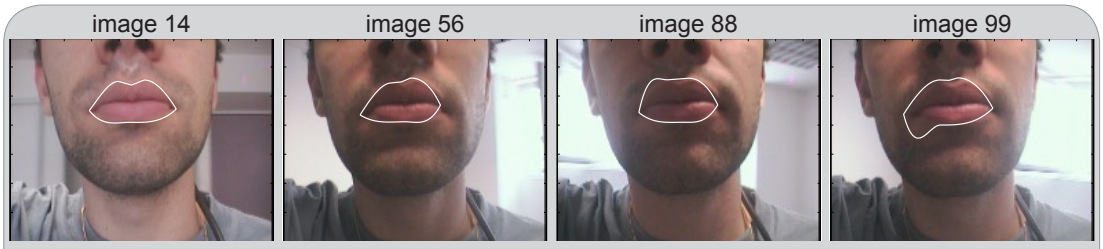


Fig. 4.31. Lorsque les contours sont peu visibles pendant un grand nombre d'images, le modèle prend peu à peu une forme aberrante.

à la courbe inférieure droite. On constate qu'il est alors impossible de modéliser correctement le contour. Si l'on impose une pente trop forte, la cubique (en trait pointillé) s'éloigne beaucoup trop de la lèvre. Pour rester proche du contour, il faut diminuer la pente, ce qui conduit à une modélisation approximative (trait plein). Ce problème pourrait être résolu en utilisant un modèle constitué de courbes dont la dérivée peut être infinie (comme, par exemple, les cubiques paramétriques). Mais l'accroissement du nombre de paramètres nécessaires peut rendre la convergence difficile. De plus, ce type de déformation des lèvres est transitoire et relativement rare. Cela ne justifie donc pas l'utilisation d'un modèle plus complexe.

La seconde limite de notre algorithme est liée à la **visibilité des contours**. Par exemple, lorsque la luminosité est trop faible , le calcul des courbes du modèle est difficile. La figure 4.31 présente une séquence pour laquelle un fort contre-jour rend les lèvres de plus en plus difficile à distinguer. Sur l'image 99, on peut constater que la courbe inférieure droite est aberrante car le contour correspondant a quasiment disparu. En général, la méthode de stabilisation du contour présentée dans la partie 4.4.2 permet de conserver une forme correcte lorsque la disparition des contours est de courte durée. Mais, sur la séquence de la figure 4.31, la bouche reste dans l'ombre pendant un grand nombre d'images et la segmentation finit par être mauvaise, malgré l'utilisation de la méthode de stabilisation. De même, lorsqu'il existe peu de différence chromatique entre les lèvres et la peau, les erreurs sont fréquentes et donnent l'impression que le modèle «ondule» autour des contours. Ce phénomène est illustré à la figure 4.32. On peut tout d'abord constater que l'initialisation n'a pu être effectuée correctement qu'à l'image 9, les

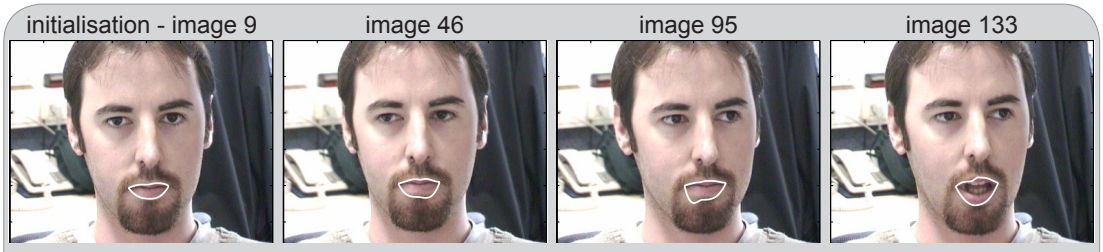


Fig. 4.32. Les erreurs sont fréquentes lorsque la couleur des lèvres est peu différente de celle de la peau.

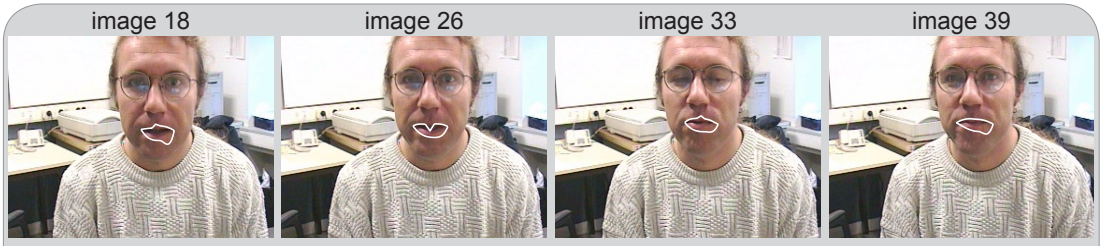


Fig. 4.33. Si les lèvres sont à la fois mal éclairées et de couleur peu différente de la peau, notre algorithme permet uniquement d'effectuer un suivi grossier de la zone de bouche.

résultats de segmentation statique obtenus sur les images précédentes étant aberrants. Sur les images suivantes, la partie inférieure du modèle suit le contour approximativement et est très souvent attirée par la limite de la zone de barbe (images 46 et 95). Enfin, des erreurs apparaissent également sur le contour de la lèvre supérieure (image 133), car son épaisseur est faible et sa limite peu marquée. Si les lèvres sont à la fois mal éclairées et de couleur peu différente de la peau, les résultats sont de très mauvaise qualité (voir figure 4.33). Dans ce cas, notre algorithme permet uniquement d'effectuer un suivi grossier de la zone de bouche.

Enfin, les **mouvements rapides** sont également susceptibles d'empêcher la segmentation précise des lèvres. En effet, l'algorithme de Lucas-Kanade ne peut suivre correctement un point lorsque l'amplitude du mouvement entre deux images consécutives est trop grande. Si l'erreur qu'il commet est trop importante, alors le recalage vers une position correcte ne peut se faire. Par exemple, sur la séquence de la figure 4.34, le locuteur ramène rapidement la tête vers la caméra. L'algorithme de Lucas-Kanade fournit alors des estimations aberrantes pour trois points sur six. Cette erreur est beaucoup trop importante pour être compensée par le recalage. De la même manière, les mouvements rapides de la lèvre inférieure peuvent également conduire à des mauvaises estimations de la position du point bas (P_6). Cependant, dans la plupart des cas, cette erreur est corrigée en quelques images. La figure 4.35 présente des séquences sur lesquelles on observe deux types de mauvais positionnement. L'ouverture rapide de la bouche (figure 4.35-a) conduit à une estimation de P_6 trop haute, et la fermeture à une estimation trop basse (figure

Fig. 4.34. Le suivi de la commissure gauche par l'algorithme de Lucas-Kanade est mauvais car le mouvement du locuteur est trop rapide. Sur l'image 102, les points caractéristiques estimés sont symbolisés par des cercles (○).

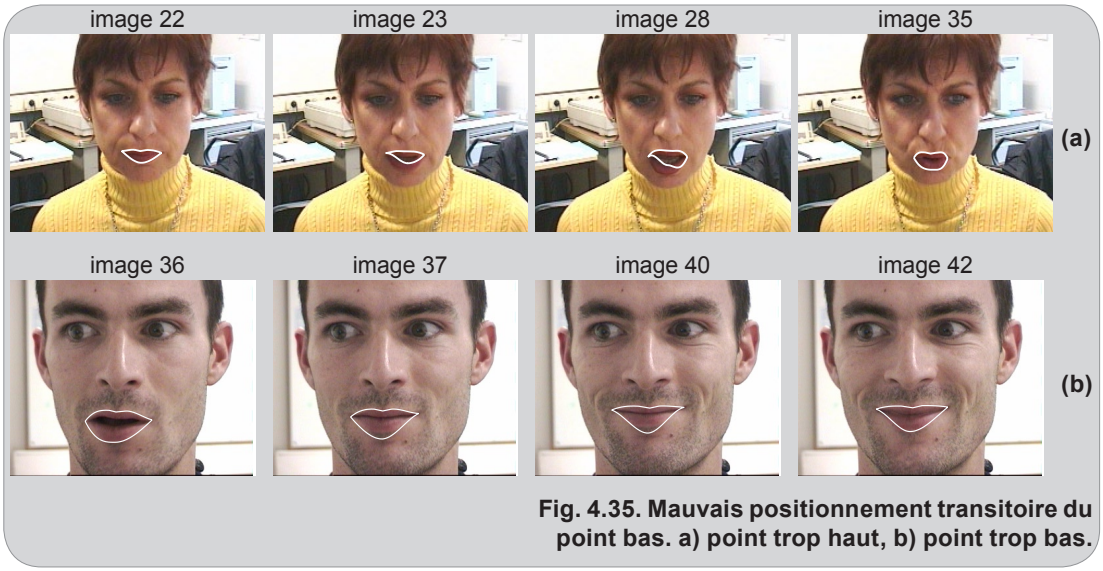


Fig. 4.35. Mauvais positionnement transitoire du point bas. a) point trop haut, b) point trop bas.

4.35-b). Dans les 2 cas, le point bas se retrouve proche d'un contour au moins aussi marqué que celui de la lèvre inférieure. Le recalage se fait donc vers ce contour, et non pas vers une position correcte. Sur la figure 4.35-a, le point est recalé sur le contour intérieur de la lèvre inférieur, et sur la figure 4.35-b, il est recalé sur la limite de la zone d'ombre située au-dessus du menton. Cependant, les erreurs de ce type sont généralement corrigées en quelques images, comme le montre les images de la figure 4.35.

4.6 Conclusion

La prise en compte d'informations temporelles améliore significativement la qualité et la rapidité de la segmentation. Dans le chapitre précédent, nous avons montré que, lors de la *segmentation statique*, l'ajustement du modèle aux contours se fait en deux étapes : **estimation de la position** et **estimation de la forme**. Pour effectuer la *segmentation dynamique*, nous avons repris la même méthodologie en incluant des informations temporelles dans chacune de ces étapes. D'une part, le suivi des points caractéristiques permet d'obtenir la position du modèle. Nous avons choisi pour cela un algorithme rapide et relativement précis : l'algorithme de Lucas-Kanade. D'autre part, comme la bouche se déforme assez peu d'une image à l'autre, nous avons facilité l'estimation de la forme en limitant l'espace de recherche des courbes optimales au voisinage des courbes du modèle précédent. Finalement, la *segmentation dynamique* peut donc se résumer à deux étapes principales : **suivi de la position** et **suivi de la forme**.

Alors que la *segmentation statique* ne pouvait s'appliquer qu'à des «formes moyennes», la *segmentation dynamique* (dont l'algorithme est résumé à la figure 4.36) permet d'extraire les contours labiaux sur la plupart des images de notre base d'essai. Le modèle utilisé est suffisamment flexible pour représenter presque toutes les formes de bouches. De plus, nous avons

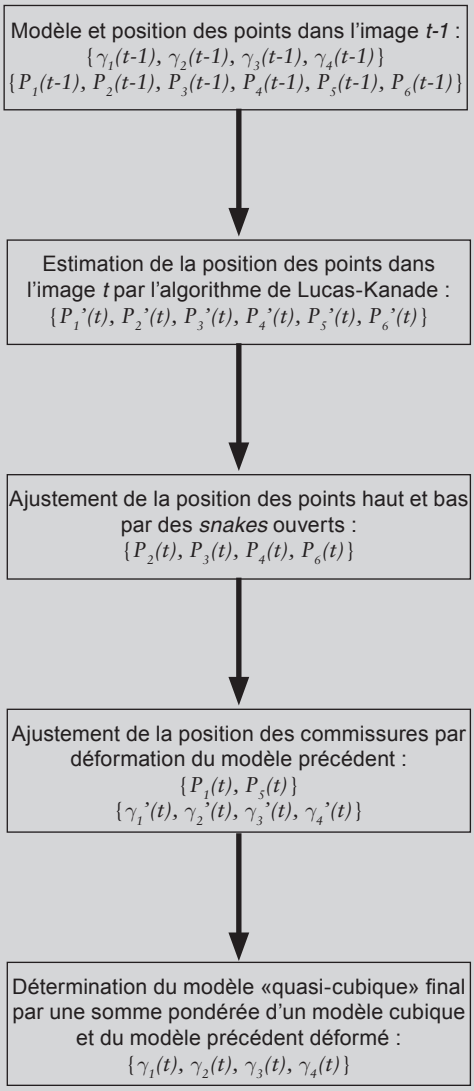


Fig. 4.36. Résumé de notre algorithme de *segmentation dynamique*.

montré que notre algorithme permet d'estimer la position des points caractéristiques avec une précision moyenne comparable à une saisie manuelle. Enfin, ses limites sont de même nature que celles de la plupart des méthodes de segmentation. Les contours doivent être suffisamment visibles et les mouvements ne doivent pas être trop rapides. Cependant, nous avons montré que, selon le type et l'étendue temporelle de la difficulté rencontrée, la méthode de recalage que nous proposons permet de compenser progressivement les erreurs.

CHAPITRE 5

Conclusion et perspectives

5.1 Conclusion

Durant de cette thèse, nous nous sommes intéressés à l'extraction des contours labiaux externes sur des séquences vidéo acquises dans des conditions naturelles avec un cadrage variable. Ces conditions de prise de vue relativement libres et les impératifs de temps réel nous ont poussés à développer un algorithme robuste et rapide. De plus, nous avons tenté d'élargir le champ des applications possibles en ajoutant des fortes contraintes de précision, indispensable notamment pour la reconnaissance d'émotion et l'identification par les lèvres.

Tout d'abord, nous avons cherché un espace couleur adapté à notre problème. Nous avons montré que la pseudo-teinte permet d'effectuer une bonne séparation des lèvres et de la peau. De plus, nous avons introduit un *gradient hybride* qui combine à la fois les informations de luminance et de chrominance, et qui facilite la localisation de la frontière supérieure de la bouche.

Ensuite, après avoir analysé les avantages et les inconvénients des méthodes classiques de segmentation labiale, nous avons opté pour une segmentation basée sur les *modèles déformables analytiques*. Cette technique est plus robuste vis-à-vis du bruit de l'image que les *méthodes de bas niveau* ou que les *contours actifs*. De plus, comme le modèle est analytique, aucun apprentissage préalable n'est nécessaire. Cependant, nous avons montré que les modèles classiques de lèvres sont trop rigides et ne fournissent souvent qu'une approximation grossière de la forme des lèvres. Nous avons donc proposé un nouveau modèle analytique composé de courbes cubiques. Celui-ci a montré qu'il était suffisamment flexible pour reproduire la plupart des formes de bouches rencontrées lors de l'élocution.

Ce modèle est construit à partir de six points caractéristiques disposés sur le contour des lèvres. Le suivi de ces points d'une image à l'autre par l'algorithme de Lucas-Kanade permet d'effectuer un positionnement rapide du modèle. Toutefois, nous avons montré que l'accumula-

tion progressive des erreurs de suivi conduit à des résultats approximatifs après quelques images. Nous avons donc proposé un algorithme de recalage (basé sur les snakes et la déformation du modèle obtenu dans l'image précédente) permettant de ramener les points vers une position correcte. Au final, nous avons montré que la compensation systématique des erreurs de suivi maintient le modèle près du contour des lèvres et permet d'envisager la segmentation de séquences très longues. De plus, la précision de ce suivi est comparable à celle d'une saisie manuelle.

Pour détecter les points caractéristiques dans la première image, nous avons introduit un nouveau type de contour actif : le *jumping snake*. Contrairement aux contours actifs classiques, ses paramètres sont faciles à régler. De plus, il est initialisé par un seul point qui peut être situé relativement loin du contour final. Le *jumping snake* permet de localiser le contour supérieur des lèvres ainsi que les points caractéristiques des lèvres. Ces points sont ensuite utilisés pour construire le modèle cubique.

Finalement, notre algorithme a permis l'extraction et le suivi des contours labiaux externes dans des conditions de prise de vue et avec des locuteurs très différents. En dehors des cas limites pour lesquels les mouvements sont trop rapides ou les contours sont peu visibles pendant un grand nombre d'images, nous avons montré qu'il est robuste vis-à-vis des variations d'éclairement, des déformations de bouche asymétriques et des rotations de tête. Dans ces conditions, les objectifs de robustesse imposés initialement par le projet *Tempovalse* ont été atteints. De plus, la précision des segmentations obtenues permet d'ores et déjà d'envisager une application de notre algorithme aux domaines de l'identification par les lèvres ou de la reconnaissance d'émotions.

5.2 Perspectives

5.2.1 Améliorations

Malgré sa robustesse et sa précision, notre algorithme n'a pas été intégré dans le système final de labiophonie du projet *TempoValse*, car nous n'avons pas pu atteindre l'objectif de traitement temps réel. Par conséquent, une première amélioration possible concerne la vitesse de traitement. Une implantation optimisée en C++, associée à l'utilisation d'une carte vidéo dédiée, devrait permettre de réduire significativement les temps de calculs. Des chercheurs du LIS spécialisés en microélectronique ont d'ailleurs déjà commencé à optimiser notre code en vue d'une réalisation logicielle et matérielle (ASIC ou FPGA). Leurs premières expérimentations en C++ permettent d'ores et déjà d'envisager un accroissement de la vitesse d'un facteur 8 ou 10.

Ensuite, notre algorithme permet uniquement d'extraire le contour labial *extérieur*. Dans l'optique d'effectuer une segmentation complète des lèvres, il serait intéressant d'appliquer la même méthodologie au contour *intérieur*. Pour cela, il faudrait commencer par trouver un gradient hybride caractérisant les contours internes des lèvres. Il faudrait ensuite déterminer un modèle déformable analytique adapté. Il est à noter que ce dernier pourrait prendre appui sur

les points P_7 et P_8, que notre algorithme permet déjà de détecter. Le réalisme du modèle *complet* de lèvres ainsi obtenu pourrait ensuite être évalué par des malentendants pratiquant la lecture labiale.

Pour le moment, le recalage est systématique. Il s'applique à tous les points et sur toutes les images. L'utilisation d'un critère quantifiant la fiabilité d'un suivi permettrait de recaler uniquement les points «problématiques». Nous avons montré que le simple calcul de la grandeur ε ne convient pas. Dans [Tomasi and Kanade, 1991], Tomasi et Kanade affirment qu'un point peut être correctement suivi si les valeurs propres de sa matrice $\boldsymbol{G}$ associée ne sont pas trop faibles. Nous pensons que la combinaison de ces valeurs propres avec ε devrait permettre d'obtenir une évaluation fiable de la qualité du suivi. Un critère de ce type a d'ailleurs déjà été proposé par Lien dans sa thèse [Lien, 1998].

Dans notre étude, nous avons montré que l'utilisation d'un jeu unique de paramètres pour le jumping snake permet de segmenter les lèvres dont la largeur est comprise entre 40 et 100 pixels. En dehors de cet intervalle, la segmentation peut se faire seulement si l'on change la valeur des paramètres. Il serait intéressant d'effectuer cette adaptation de manière automatique. Pour cela, il serait possible d'utiliser la taille du visage (les algorithmes de détection de visage sont nombreux et fiables) pour calculer la largeur approximative des lèvres, ce qui permettrait d'obtenir un jeu de paramètres adapté.

Pour le moment, les paramètres du modèle peuvent prendre n'importe quelle valeur. Une étude statistique permettrait de restreindre leurs domaines de variation. De cette manière, la convergence pourrait s'effectuer plus rapidement et les formes de bouche aberrantes pourraient être écartées. Grâce à la robustesse et à la précision de notre algorithme, cette étude statistique pourrait être menée assez facilement. Au lieu de placer manuellement le modèle sur la bouche, un expert validerait simplement la segmentation effectuée automatiquement. Si la segmentation est mauvaise, l'expert aurait le choix de passer à l'image suivante ou d'adapter lui-même le modèle sur le contour des lèvres. D'après nos estimations, cette technique permettrait d'étiqueter plusieurs milliers d'images en quelques heures.

5.2.2 Applications

Nous avons d'ores et déjà commencé une coopération avec Zakia Hammal, une doctorante du LIS, sur un projet de reconnaissance automatique d'expressions faciales [Hammal *et al.*, 2003]. Le système actuel utilise des *modèles déformables analytiques* pour détecter la position et la forme des sourcils, des yeux et de la bouche (voir figure 5.1). Une classification des paramètres faciaux extraits permettra ensuite d'effectuer une reconnaissance d'expressions.

Ensuite, les six points caractéristiques que nous détectons sont également utilisés dans la norme MPEG-4 (voir figure 1.6). De plus, les points intermédiaires 8.5, 8.6, 8.7 et 8.8 peuvent être obtenus sur les courbes de notre modèle, à mi-distance des points caractéristiques principaux. Théoriquement, il est donc possible d'intégrer notre algorithme dans une chaîne de communication basée sur le standard MPEG-4.

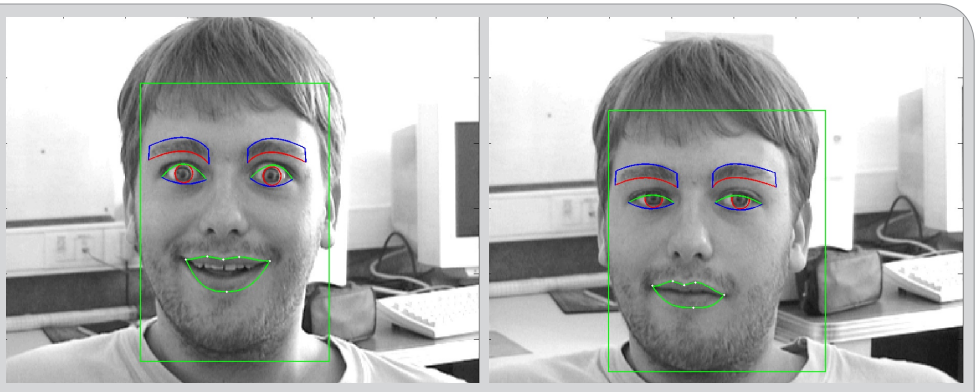


Fig. 5.1. Notre algorithme a déjà été intégré à un système de reconnaissance d'expression, d'après [Hammal *et al.*, 2003].

A ce jour, le domaine de l'identification par les lèvres est encore relativement vierge. Les quelques auteurs qui s'y sont intéressés soulignent tous l'importance fondamentale de la précision de la segmentation [Brand, 2001][Chibelushi, 1997]. Etant donné le réalisme des segmentations que nous obtenons, il serait intéressant d'évaluer l'apport de notre algorithme pour l'identification du locuteur.

Références bibliographiques

[Amini *et al.*, 1990] A. Amini, T. Weymouth, and R. Jain. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):855--867, 1990.

[Anandan, 1989] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. In *International Journal on Computer Vision*, vol. 2, pp. 283-310, 1989.

[AVSP, 1998] Procedings of *Auditory-Visual Speech Processing*, Terrigal, Australia, 1998.

[Bailly, 2001] G. Bailly. Audiovisual speech synthesis. In Taylor, P., editor, *ETRW on Speech Synthesis*, Perthshire - Scotland, 2001.

[Baron et al., 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.

[Belhumeur *et al.*, 1997] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. sherfaces: Recognition using specic linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711--720, July 1997.

[Benoît *et al.*, 1992] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of French visemes for visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 485--501.

[Berger and Mohr, 1990] M.O. Berger and R. Mohr. Towards Autonomy in Active Contour Models", In *Proc. 10th international Conference on Pattern Recognition (ICPR'90)*, Atlantic City, June 1990.

[Beskow *et al.*, 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 *Proc. of Eurospeech '97*, Rhodos, Greece, 1997.

[Botino, 2002] A. Botino. Real time head and facial features tracking from uncalibrated monocular views. In *Proc. 5th Asian Conference on Computer Vision (ACCV'02)*, Melbourne,

Australia, 23-25 January 2002.

[Brand, 2001] J.D. Brand. Visual speech for speaker recognition and robust face detection. PhD thesis, University of Wales, UK, 2001

[Caselles *et al*., 1995] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *Proc. 5th IEEE Int. Conf. on Computer Vision (ICCV)*, pp 694--699, Cambridge, Massuchusetts, 1995.

[Chakraborty *et al*., 1994] A. Chakraborty, L. H. Staib, and J. S. Duncan. Deformable boundary finding influenced by region homogenity. In *Proc. IEEE Conf. on Computer Vision and Pattren Recognition (CVPR)*, pp 624--627, Seattle, WA, June 1994.

[Chan *et al*., 1998] M.T. Chan, Y. Zhang, and T.S. Huang. Real-time lip tracking and bimodal continuous speech recognition. In *Proc. 2nd MMSP*, pp 65--70, Los Angeles, CA, USA, December 1998.

[Chen and Rao, 1998] T. Chen. and R.R. Rao. Audio-Visual Integration in Multimodal Communication. In *IEEE Special Issue on Multimedia Signal Processing*, May 1998.

[Chibelushi, 1997] C. Chibelushi. Automatic Audio-Visual Person Recognition. PhD thesis, University of Wales, Swansea, 1997.

[Chiou and Hwang, 1997] G. I. Chiou and J.-N. Hwang. Lipreading from color video. In *Trans. on Image Processing*, 6(8):1192-1195, August 1997.

[Cohen, 1991] L. Cohen. Note on active contour models and ballons. In *Proc. CVGIP: Image Understanding*, 53(2):211-218, March 1991.

[Cohen and Cohen, 1993] L. D. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-D and 3-D images. In *IEEE Transactions Pattern analysis and Machine Intelligence*. 15(11):1131-1147, November 1993.

[Cohen *et al*., 1995] M.M. Cohen, R.L. Walker, and D.W. Massaro. Perception of synthetic visual speech. In D. Stork & M. Hennecke M editors, *Speechreading by Humans and Machines*: NATO ASI Series, Springer, 153-168, 1995.

[Coianiz *et al*., 1995] T. Coianiz, L. Torresani, and B. Caprile. 2D Deformable Models for Visual Speech Analysis. In *NATO Advanced Study Institute: Speech reading by Man and Machine*, 1995.

[Cootes and Taylor, 1992] T. F. Cootes and C. J. Taylor. Active Shape Models - «Smart Snakes». In *Proc. British Machine Vision Conference,* pp 266-275, Springer-Verlag, 1992.

[Cootes *et al*., 1993] T.F. Cootes, C.J. Taylor, A. Lanitis, D. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proc. ICCV*, pp 242-246, 1993.

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

[Cootes *et al*., 1998] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, pp 484–498, June 1998.

[Cootes *et al.*, 1999] T.F. Cootes, G. J. Edwards and C. J. Taylor. Comparing Active Shape Models with Active Appearance Models. In *Proc. British Machine Vision Conference*, Vol. 1, pp. 173-182, 1999.

[Cosatto and Graf, 1998] E. Cosatto and H.P. Graf. Samplebased of photo-realistic talking heads. In *Computer Animation*. pp 103-110, Philadelphia, Pennsylvania, 1998.

[Daubias, 2002] P. Daubias. Modèles *a posteriori* de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle. Thèse de doctorat, Université du Maine, 2002.

[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.

[Dodd and Campbell, 1987] B. Dodd and R. Campbell. Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Associates, London, 1987.

[Easton and Basala, 1982] R.D. Easton and M. Basala. Perceptual dominance during lipreading. *Perception and Psychophysics*, vol. 32, pp 562-570, 1982.

[Ekman and Friesen, 1978] P. Ekman and W.V Friesen. Facial action coding system. Palo Alto: Consulting Psychologist Press.

[Erber, 1969] N.P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12: 423-425.

[Essa, 1995] I.A. Essa. Analysis, Interpretation and Synthesis of Facial Expressions. *MIT Media Laboratory*, Perceptual Computing Technical Report 303, Ph.D. dissertation, February 1995.

[Essa and Pentland, 1994] I.A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proc. of Computer Vision and Pattern Recognition (CVPR 94)*, pp 76–83, 1994.

[Etemad and Chellapa, 1997] K. Etemad and R. Chellappa. Discriminant Analysis fo Recognition of Human Face Images. In *Proc. AVBPA, Lecture Notes in Computer Science 1206*, pages 127-142, 1997.

[Eveno *et al.*, 2002] N. Eveno, A. Caplier and P.Y. Coulon. A Parametric Model for Realistic Lip Segmentation. 7th *International Conference on Control, Automation, Robotics and Vision (ICARCV'02)*, Singapore, December 2002.

[Faruquie *et al.*, 2000] T. A. Faruquie, A. Majumdar, N. Rajput, L. V. Subramaniam. Large Vocabulary Audio-Visual Speech Recognition using Active Shape Models. In *Proc. International Conference on Pattern Recognition (ICPR 2000)*, Barcelona, Spain, September 3-8, 2000.

[Finn and Montgomery, 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.

[Fisher, 1968] C. Fisher. Confusions among visually perceived consonants. *Journal of Speech And Hearing Research*, vol. 11, pp. 796-804, 1968.

[Fleet and Jepson, 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.

[Fua and Brechbuhler, 1996] P. Fua and C. Brechbuhler. Imposing hard constraint on sift snakes. In *Proc. European Conf. Computer Vision '96*, vol. 2, pp 495--506, 1996.

[Gao *et al.*, 1998] J. Gao, A. Kosaka and A. Kak. A deformable model for human organ extraction. *International Conference on Image Processing*, 1998.

[Geiger *et al.*, 1993] D. Geiger, A. Gupta, L. Costa and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3)294--302, March 1993.

[Goldshen, 1993] A.J. Goldschen. Continuous Automatic Speech Recognition by Lipreading. Ph.D. dissertation, George Washington University, September 1993.

[Gordan *et al.*, 2002] M. Gordan, C. Kotropoulos, A. Georgakis, I. Pitas. A new fuzzy C-means based segmentation strategy. Application to lip region identification. *IEEE-TTTC International Conference on Automation, Quality and Testing, Robotics*, Cluj-Napoca, Romania, May 23-25, 2002

[Hammal *et al.*, 2003] Z. Hammal, N. Eveno, A. Caplier and P.Y. Coulon. Extraction réaliste des traits caractéristiques du visage à l'aide de modèles paramétriques adaptés. *Colloque GRETSI sur le traitement du signal et de images (GRETSI'03)*, Paris, France, 2003.

[Heeger, 1988] D. J. Heeger. Optical flow using spatiotemporal filters. In *International Journal on Computer Vision*, vol. 1, pp. 279-302, 1988.

[Hennecke *et al.*, 1994] M.E. Hennecke, K.V. Prasad and D.G. Stork. Using deformable templates to infer visual speech dynamics. *28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 1994

[Hildreth, 1984] E. C. Hildreth. The measurement of visual motion. Cambridge, MA: MIT Press, 1984.

[Himer *et al.*, 1991] W. Himer, F. Schneider, G. Kost and H. Heimann. Computer-Based Analysis of Facial Action: A New Approach. *Journal of Psychophysiology*, Vol. 5, No. 2, pp. 189-195, 1991.

[Horbelt and Dugelay, 1995] S. Horbelt and J. L. Dugelay. Active contours for lipreading - combining snakes with templates. *GRETSI symposium on Signal and Image Processing*, France, 1995.

[Horn and Schunk, 1981] B.K.P Horn and B.G. Schunck. Determining optical flow. *AI* 17, pp. 185-204, 1981.

[Hsu *et al.*, 2002] R.-L. Hsu, M. Abdel-Mottaleb and A. K. Jain. Face detection in color images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706, May 2002.

[Hulbert and Poggio, 1998] A. Hulbert and T. Poggio. Synthesizing a Color Algorithm From Examples. *Science*, Vol 239, pp 482-485, 1998.

[ITU, 1995] ITU-T SG15 WP15/1, Draft Recommendation H.263 (video coding for low bitrate

communications), Document LBC-95-251, October 1995.

[Jourlin *et al.*, 1997] P. Jourlin, J. Luettin, D. Genoud and H. Wassner. Acoustic Labial Speaker Verification. *Proc AVBPA, Lecture Notes in Computer Science 1206*, pages 319-334, 1997.

[Kähler *et al.*, 2002] K. Kähler, J. Haber, H. Yamauchi, H.-P. Seidel. Head shop: Generating animated head models with anatomical structure. *In. Proc. ACM SIGGRAPH Symposium on Computer Animation (SCA'2002)*, pp. 55-64, 21-22 July 2002.

[Kass *et al.*, 1987] M. Kass, A. Witkin and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pp 321-331, 1987.

[Koga, 1981] T. Koga. Motion compensated interframe coding for video conferencing. *National telecommunication conference*, New Orleans, November 1981.

[Lallouache, 1991] T. Lallouache. Un poste Visage-Parole. Acquisition and traitement automatique des contours des lèvres. Thèse de doctorat, Institut National Polytechnique de grenoble, 1991.

[Lavagetto, 1995] Lavagetto, F., "Converting speech into lip movements: A multimedia telephone for hard of hearing people," IEEE Transactions on Rehabilitation Engineering, vol. 3, no. 1, pp. 1-14, March 1995

[Leroy, 1996] B. Leroy. Modèles déformables et modèles de déformation appliqués à la reconnaissance de visages. Thèse de doctorat, Université Paris IX-Dauphine, Juin 1996.

[Leymarie and Levine, 1993] F. Leymarie and M. Levine. Tracking deformable objects in the plane using an active contour model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):617--634, June 1993

[Li *et al.*, 1993] Haibo Li, Pertti Roivainen and Robert Forchheimer. 3-D Motion Estimation in Model-Based Facial Image Coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 6, June 1993 p. 545 - 555.

[Lien, 1998] J.J. Lien. Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity. PhD thesis, Carnegie Mellon University, Pittsburg, april 1998.

[Lievin and Luthon, 1999] M. Lievin and F. Luthon. Unsupervised Lip Segmentation under Natural Conditions. In *ICASSP'99*, Phoenix, Arizona, pp. 3065–3068, 1999.

[Liew *et al.*, 1999] W.C. Alan Liew, K.L. Sum, S.H. Leung, W.H. Lau. Fuzzy segmentation of lip image using cluster analysis. *European Conference on Speech Communication and Technology (EUROSPEECH'99)*, Hungary, 1999

[Liew *et al.*, 2000] A.W.C. Liew, S.H. Leung, and W.H. Lau. Lip contour extraction using a deformable model. *Int. Conf. on Image Processing (ICIP'00)*, Vancouver, Canada, 2000.

[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 *et al.*, 1999] S. Lucey, S. Sridharan and V. Chandran. Chromatic lip tracking using a connectivity based fuzzy thresholding technique, In *ISSPA'99*, 1999

[Lucey *et al.*, 2000] S. Lucey, S. Sridharan, and V. Chandran. Face and lip tracking using chromatic based AVQ. Technical Report, 2000.

[Luettin *et al.*, 1995] J. Luettin, N.A. Tracker, S.W. Beet. Active Shape Models for Visual Speech Feature Extraction. Electronic System Group Report N°95/44, University of Sheffield, UK, 1995.

[Luettin *et al.*, 1996] J. Luettin, N. Thacker and S. Beet. Visual speech recognition using active shape models and hidden Markov models. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96)*, vol. 2, pp. 817--820, 1996.

[Luettin *et al.*, 1996+] J. Luettin, N.A. Thacker, and S.W. Beet. Speaker Identification by Lipreading. In *Proc ICSLP*, pages 62-64, 1996.

[Luettin, 1997] J. Luettin. Visual Speech and Speaker Recognition. PhD thesis, University of Sheeld, May 1997.

[Lyons *et al.*, 2003] Michael J. Lyons, Michael Haehnel and Nobuji Tetsutani. Designing, Playing, and Performing with a Vision-Based Mouth Interface. In *Proc. Conference on New Interfaces for Musical Expression (NIME-03)*, pp. 116-121, 2003.

[Mase and Pentland, 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 *et al.*, 1998] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J.A. Bangham. Lipreading using shape, shading and scale. In *Proc. Auditory-Visual Speech Processing (AVSP)*, pp. 73-78, Australia, December 1998.

[Matthews *et al.*, 1998+] I. Matthews, J.A. Bangham, R. Harvey, and S. Cox, A comparison of active shape model and scale decomposition based features for visual speech recognition. *LNCS*, 1407: 514--528, 1998.

[McGurk and McDonald, 1976] H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature*, pp. 746-748, December 1976.

[Morishima *et al.*, 1989] S. Morishima, K. Aizawa and H. Harashima. An intelligent facial image coding driven by speech and phoneme. In *Proc. IEEE ICASSP*, p. 1795, Glasgow, UK, 1989.

[Morishima, 1995] S. Morishima. Emotion model. *International Workshop on Automatic Face and Gesture Recognition*, Zurich, pp. 284-289, 1995.

[MPEG, 1997] MPEG-N1902, Text for CD 14496-2 Video, *ISO/IEC JTC1/SC29/WG11 N1886*, MPEG97/November 1997.

[Nagel, 1987] N.-H. Nagel. On the estimation of optical flow : relations between different approaches and some new results. *AI* 33, pp. 299-324, 1987.

[Nefian *et al.*, 2002] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. A coupled HMM for audio-visual speech recognition. In *Proc. ICASSP*, volume II, pages 2013-2016, Orlando, FL, USA, May 2002.

[Neuenschwander *et al.*, 1997] W. Neuenschwander, P. Fua, L. Iverson, G. Szekely, and O. Kuebler. *Ziplock Snakes. International Journal of Computer Vision*, 26(3):191--201, December

1997.

[Nishida, 1986] S. Nishida. Speech recognition enhancement by lip information. *ACM SIGCHI Bulletin*, 17 (4), 198-204, 1986.

[Odisio and Bailly, 2003] M. Odisio and G. Bailly. Shape and appearance models of talking faces for model-based tracking. *Auditory-visual Speech Processing Workshop (AVSP'03)*, France, 2003

[Oliver *et al*., 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.

[Pantic *et al*., 2001] M. Pantic, M. Tomc and L.J.M. Rothkrantz. A hybrid approach to mouth features detection. In *Proc. IEEE Int'l Conf. on System, Man and Cybernetics*, pp. 1188-1193, Tucson, Arizona, USA, October 2001.

[Parke, 1982] F.I. Parke.Parameterised models for facial animation. *IEEE Computer Graphics and Applications*, vol. 12, pp. 61-68, November 1982.

[Patterson *et al*., 2003] 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. accepted for publication by the *EURASIP Journal on Applied Signal Processing*. To Appear in 2003.

[Petajan, 1984] Eric Petajan. Automatic Lipreading to Enhance Speech Recognition. PhD thesis, University of Illinois at Urbana-Champaign, 1984.

[Petajan *et al*., 1988] E.D. Petajan, B. Bischoff, D. Bodoff and N.M. Brooke. An improved automatic lipreading system to enhance speech recognition. *CHI88*, pages 19-25, 1988.

[Potamianos *et al*., 1997] G. Potamianos, E. Cosatto, H.P. Graf and D.B.Roe. Speaker independent audio-visual database for bimodal ASR. In *Proc. of the European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, Spet.1997.

[Potamianos *et al*., 2004] G. Potamianos, C. Neti, J. Luettin and I. Matthews. Audiovisual automatic speech recognition: an overview. Audiovisual Speech Processing. E. Bateson, G. Bailly and P. Perrier editors, MIT Press, 2004.

[Rabiner and Juang, 1993] L.R. Rabiner, and B.H. Juang. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.

[Radeva and Marti, 1995] P. Radeva and E. Marti. Facial features segmentation by model-based snakes. *International Conference on Computer Analysis of Images and Patterns*, 1995.

[Radeva *et al*., 1995] P. Radeva, J. Serrat, and E. Marti. A snake for model-based segmentation. *International Conference on Computer Vision*, 1995.

[Rao and Mersereau, 1995] R. Rao and R. Mersereau. On merging hidden Markov models with deformable templates. *ICIP 95*, Washington D.C., 1995.

[Revéret *et al*., 1997] L. Revéret, F. Garcia, C. Benoit, E. Vatikiotis-Bateson (1997), "An hybrid approach to orientationfree liptracking", AVSP'97, 117-120.

[Rydfalk, 1987] Rydfalk, M., "CANDIDE: A Parameterized face," Report LiTH-ISY-I-0866, Linkoping University, October 1987.

[Stork *et al*., 1992] Stork, D. G., Wolff, G., and Levine E., "Neural network lipreading system for improved speech recognition," Intl. Joint Conf. on Neural Networks, pp. 285-295, 1992.

[Sumby and Pollack, 1954] Sumby WH & Pollack I (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26: 212-215.

[Terzopoulos and Waters, 1990] D. Terzopoulos and K. Waters. Physically-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, **1**: p. 73--80.

[Terzopoulos and Waters, 1993] D. Terzopoulos, K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models". *IEEE Trans Pattern Analysis and Machine Intelligence*, 15(6), pp. 569-579, June 1993.

[Tian *et al*., 2000] Y. Tian, T. Kanade, J. Cohn, "Robust Lip Tracking by Combining Shape, Color and Motion", 4th Asian Conference on Computer Vision (ACCV'00), January, 2000.

[Tsapatsoulis *et al*., 2000] N. Tsapatsoulis, Y. Avrithis and S.Kollias, "Efficient Face Detection for Multimedia Applications", ICIP00, TA07.11, Vancouver, Canada, September 2000.

[Tomasi and Kanade, 1991] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report CMU-CS-91-132, *Carnegie Mellon University*, 1991.

[Uras *et al*., 1988] S. Uras, F. Girosi, A. Verri and V. Torre. A computational approach to motion perception. In *Biol. Cybern.* 60, pp 79-97, 1988.

[Vignoli and Braccini, 1999] Vignoli, F. and C. Braccini. A textspeech synchronization technique with applications to talking heads. In *AVSP'99*, Santa Cruz, California, USA, 1999.

[Vezhnevets, 2002] V. Vezhnevets. Face and facial feature tracking for natural Human-Computer Interface. *Graphicon'2002*, September 16 - September 21, 2002, Nizhny Novgorod, Russia.

[Wark and Sridharan, 1998] T. Wark and S. Sridharan. A syntatic approach to automatic lip feature extraction for speaker identification. *ICASSP' 98*, pp. 3693–3696.

[Welsh *et al*., 1990] W.J. Welsh, A.D. Simons, R.A. Hutchinson, and S. Searby. Synthetic face generation for enhancing a user interface. In *Proc. Image'Com Conf*., pp. 177-182, Bordeaux, November 1990.

[Wolberg, 1990] G. Wolberg. Digital Image Warping. IEEE Computer Society Press, 1990.

[Xu and Prince, 1998] C. Xu and J.L. Prince. Snakes, shapes and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359--369, Mars 1998.

[Yacoob and Davis, 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.,* 35 (4) (1993) 172}181.

[Yang *et al*., 1997] J. Yang, W. Lu and A. Waibel. Skin-color modeling and adaptation. Technical Report CMU-CS-97-146, School of computer Science, Carnegie Mellon University, 1997.

[Yang and Waibel, 1996] J. Yang and A. Waibel. A Real-Time Face Tracker. In *Proc. of WACV'96*, pp. 142–147, Sarasota, USA, 1996.

[Yuhas *et al*., 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, Nov. 1989.

[Yuille *et al*., 1992] A. Yuille, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *Int'l Journal of Computer Vision*, vol. 8, no. 2, pp. 99-111, 1992.

[Zarit *et al*., 1998] B.D. Zarit, B.J. Super, and F.K.H. Quek. Comparison of five color models in skin pixel classification. In *Proc. of the IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 1998.

[Zhang, 1997] L. Zhang. Estimation of the mouth features using deformable templates. *IEEE International Conference on Image Processing (ICIP 97)*, Vol. III, pp. 328-331, Santa Barbara, CA., October 1997.

[Zhang and Mersereau, 2000] X. Zhang and R.M. Mersereau. Lip Feature Extraction Towards an Automatic Speechreading System. *ICIP*, 2000.

Publications

[Eveno *et al.*, 2004] N. Eveno, A. Caplier, and P-Y Coulon. Automatic and Accurate Lip Tracking. *IEEE Transaction on circuits and video technology*, parution prévue en mai 2004.

[Eveno *et al.*, 2003] N. Eveno, A. Caplier, and P-Y Coulon. Jumping Snakes and Parametric Model for Lip Segmentation. *International Conference on Image Processing* (*ICIP'03*), Barcelona, Spain, September 2003.

[Hammal *et al.*, 2003] Z. Hammal, N. Eveno, A.Caplier, and P-Y Coulon. Extraction réaliste des traits caractéristiques du visage à l'aide de modèles paramétriques adaptés. *Colloque GRETSI sur le traitement du Signal et des Images (GRETSI'03*), Paris, France, September 2003.

[Eveno *et al.*, 2002] N. Eveno, A. Caplier, and P-Y Coulon. A Parametric Model for Realistic Lip Segmentation. *International Conference On Robotics Control and Vision (ICARCV'02)*, Singapore, December 2002.

[Eveno *et al.*, 2002] N. Eveno, A. Caplier, and P-Y Coulon. Keypoints Based Segmentation of Lips. *International Conference On Multimedia and Expo (ICME'02)*, Lausanne, Switzerland, September 2002

[Delmas *et al.*, 2002] P. Delmas, N. Eveno, and M. Lievin. Towards Robust Lip Tracking. *International Conference on Pattern Recognition (ICPR'02)*, Québec City, Canada, August 2002.

[Eveno *et al.*, 2001] N. Eveno, A. Caplier, and P-Y Coulon. A New Color Transformation for Lip Segmentation", *Multimedia Signal Processing (MMSP'01),* Canne, France, October 2001.

[Eveno *et al.*, 2001] N. Eveno, P. Delmas, and P-Y Coulon. Towards automatic lip tracking. *Image and Vision Computing New Zealand (ICVNZ'01),* Dunedin, New Zealand, November 2001.

[Eveno *et al.*, 2001] N. Eveno, P. Delmas, and P-Y Coulon. Vers l'Extraction Autotmatique des Lèvres d'un Visage Parlant. *Colloque GRETSI sur le traitement du Signal et des Images (GRETSI'01*), Toulouse, France, September 2001.

La segmentation des lèvres est une étape essentielle pour de nombreux systèmes multimedia tels que la vidéoconférence, la lecture labiale ou les systèmes de communication bas débit. Au cours de cette thèse, nous avons développé un algorithme quasi automatique, précis et robuste de segmentation de lèvres dans des séquences vidéo. Dans un premier temps, le contour supérieur de la bouche ainsi que plusieurs points caractéristiques sont détectés dans l'image initiale en utilisant un nouveau type de contour actif nommé "jumping snake". Contrairement aux snakes classiques, le jumping snake est peu sensible à l'initialisation et la détermination de ses paramètres est simple et intuitive. Pour la segmentation proprement dite, nous introduisons un modèle analytique très flexible composé de quelques courbes cubiques. L'intérêt de ce modèle réside avant tout dans sa grande flexibilité qui permet de rendre compte de manière réaliste d'un très large panel de formes possibles pour la bouche. Dans les images suivantes, la segmentation est réalisée en utilisant un suivi temporel des points caractéristiques et des paramètres du modèle. De plus, nous proposons un algorithme de recalage permettant de compenser efficacement les erreurs de suivi. Finalement, nous montrons que notre algorithme permet de suivre les points caractéristiques avec une précision comparable à celle d'une saisie manuelle.

Lip segmentation by using an analytical deformable model.

Lip segmentation is an essential stage in many multimedia systems such as videoconferencing, lip reading, or low bit rate coding communication systems. In this paper, we propose an accurate and robust quasi automatic lip segmentation algorithm. First, the upper mouth boundary and several characteristic points are detected in the first frame by using a new kind of active contour : the "jumping snake". Unlike classic snakes, it can be initialized far from the final edge and the adjustment of its parameters is easy and intuitive. Then, to achieve the segmentation we propose a parametric model composed of several cubic curves. Its high flexibility enables accurate lip contour extraction even in the challenging case of very asymmetric mouth. Compared to existing models, it brings a significant accuracy and realism improvement. The segmentation in the following frames is achieved by using an interframe tracking of the keypoints and the model parameters. However, we show that, with a usual tracking algorithm, the keypoints positions become unreliable after a few frames. We therefore propose an adjustment process that enables an accurate tracking even after hundreds of frames. Finally, we show that the mean keypoints tracking errors of our algorithm are comparable to manual points selection errors.

Mots-clefs : segmentation, lèvres, modèle, contours actifs, couleur, suivi.

Keywords : segmentation, lips, model, active contours, color, tracking.

Laboratoire des Images et Signaux
46 avenue Felix Viallet
38031 Grenoble Cedex