



INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque

THESE EN COTUTELLE INTERNATIONALE

pour obtenir le grade de

**DOCTEUR DE L'Institut polytechnique de Grenoble
et
de l'Université Laval**

Spécialité : Signal, Image, Parole, Télécoms

préparée au laboratoire GIPSA-lab/DIS

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

et au laboratoire de Vision et Systèmes Numériques de l'Université Laval

présentée et soutenue publiquement

par

Christian Bouvier

Le 02/02/2010

SEGMENTATION REGION-CONTOUR DES CONTOURS DES LEVRES

Sous la direction de M. **Pierre-Yves Coulon**
Sous la co-direction de M. **Xavier Maldaque**

JURY

Mme. Alice Caplier
M. Franck Luthon
M. Maurice Milgram
M. Pierre-Yves Coulon
M. Xavier Maldaque
M. Robert Bergevin
M. Christophe Blanc

, Président
, Rapporteur
, Rapporteur
, Directeur de thèse
, Co-directeur de thèse
, Examinateur
, Examinateur

Résumé

La thèse présentée a été effectuée en cotutelle entre l’Institut Polytechnique de Grenoble et l’université Laval à Québec. Les travaux ont impliqué les laboratoires universitaire GIPSA-lab à Grenoble et le Laboratoire de Vision et Systèmes Numériques de l’Université Laval (LVSN). Notre étude porte sur la segmentation des contours internes et externes des lèvres. L’objectif visé dans notre étude est de proposer un ensemble de méthodes permettant de modéliser précisément la zone de la bouche avec la meilleure robustesse possible. Par robustesse, nous entendons obtenir une méthode fiable qui ne nécessite pas de réglage de paramètres et qui permette une segmentation fidèle des contours externes et internes de la bouche.

Dans un premier temps, une approche combinée région-contour est introduite dans le but d’obtenir une segmentation multi-locuteur de la bouche sur des images de visage en couleurs. Nous décrivons une approche par décorrélation permettant d’augmenter le contraste entre la peau et les lèvres sur des images en couleurs ainsi qu’une étude sur les gradients multi-échelles pour améliorer la robustesse de la modélisation des contours de la bouche. Ensuite nous présentons notre méthode de localisation et de segmentation région-contour de la bouche sur des images de visage en couleurs.

Dans un second temps nous nous sommes intéressés à la détection de l’état de la bouche. L’étape de détection de l’état de la bouche est nécessaire à la modélisation de la région interne qui présente une grande variabilité de forme et de texture. Une approche bio-inspirée, basée sur un modèle de rétine et de cortex visuel conduisant au calcul d’un spectre Log-polaire, a été développée pour modéliser la zone de la bouche. Ces spectres sont, ensuite, utilisés pour entraîner un réseau *SVM* destiné à identifier l’état de la bouche.

D’autre part, de nombreux auteurs ont suggéré d’utiliser la modalité infrarouge en analyse faciale. Le LVSN possédant une solide expertise dans le domaine de la vision infrarouge, une étude sur la pertinence de la modalité infrarouge dans le cadre de la segmentation des lèvres est proposée.

Après avoir localisé la bouche et identifié l’état de la bouche, nous nous intéressons alors à la segmentation des contours externes et internes de la bouche. Un modèle polynomial de contour externe, dont la complexité sera automatiquement adaptée en fonction de la bouche

traitée, est présenté. L'aspect de la zone interne de la bouche pouvant varier très rapidement, l'extraction du contour interne est particulièrement difficile. Nous proposons de traiter ce problème par une méthode de classification non-supervisée pour sélectionner les régions internes de la bouche. La méthode de modélisation de contour par un modèle polynomial est par la suite appliquée pour extraire le contour interne de la bouche.

Enfin, une analyse quantitative de la performance globale de l'extraction des contours internes et externes est réalisée par comparaison avec des vérités-terrain.

Abstract

This report presents the thesis that has been jointly conducted at the Grenoble Institute of Technology in France and at the Laval University in Canada. The work involves the GIPSA-lab in Grenoble and the CVSL of the Laval University. The aim of the thesis is to propose a set of robust methods to segment the lips outer and inner contours. In terms of robustness, we intend to propose a reliable lips contours segmentation that does not require the tuning of parameters.

In the first chapter a state of the art of mouth analysis is given.

In the second chapter of this report, we will introduce our “region-contour” based approach to segment a binary mask of the lips on static face color images. First, we will describe the decorrelation-stretch algorithm we use to enhance the contrast between lips pixels and skin pixels and then a multi-scale gradients approach to contour modeling. The last part of the chapter will focus on the segmentation of a lips binary mask by automatic thresholding of a specific chromatic component.

In chapter three we will be interested in the mouth state identification problem. The knowledge of mouth state is critical if one is to propose a robust segmentation of the lips internal contours. A bio-inspired approach based on retina and visual cortex models has been developed to compute a scale invariant mouth description: the log-polar spectrum. Log-polar spectrums, computed on a manually classified mouth images database, are used to train a *SVM* network. The goal of the network is to classify unknown mouth images in 2 clusters: open mouth and closed mouth.

Recently, infrared based approaches have become popular in face analysis, especially for face recognition problems. Infrared thermography is an area of extensive expertise at the CVSL. In order to enhance the robustness of the lips contours segmentation, we studied the potential of the information given by infrared face images. A combined visible/infrared face image database has been constructed for that purpose. Chapter 4 describes the construction of the combined visible/infrared database and the statistical study of the skin/lips contrast on infrared face images.

The last chapter deals with the outer and inner lips contours segmentation. The mouth outer contour is processed first. Using the lips binary mask given by the algorithm described in

chapter 2, we have developed an adaptive contour model for the external mouth contour. The contour will be described by a polynomial curve. The complexity of the curve will be adapted using color and multi-scale gradients information. For the inner lips contour, we proposed an unsupervised classification method to segment the inner areas of the mouth. This gives us a binary mask of the inner areas of the mouth. Finally, given the mask of the inner areas of the mouth, the contour is extracted by using the same method that has been applied on the outer contour. The last section of chapter 5 will deal with the performance evaluation of our segmentation algorithms. An original approach for contour comparison, based on Fourier descriptors, has been developed for that purpose.

Remerciements

Tout d'abord, je souhaite remercier Pierre-Yves Coulon pour sa disponibilité et ses conseils tout au long de ma thèse. Je tiens également à remercier Xavier Maldague pour son aide et ses conseils dans le cadre de la cotutelle de thèse à l'université Laval.

Je tiens également à remercier les nombreuses personnes avec lesquelles j'ai collaboré pendant les 4 années et demie que j'ai passées au Gipsa-lab et au LVSN: Alexandre Benoit, Pierre Gacon, Matthieu Klein, Jean-Marc Piau, Alice Caplier, Abdel Hakim Bendada.

Merci également à Jean-Marc Sache, Hervé Colasuonno et Denis Ouellet pour leur soutien technique.

Enfin, je tiens à remercier mes proches. Merci à ma sœur pour son soutien. Merci à mes grands-parents pour leur aide, j'ai une pensée émue en particulier pour mon grand-père décédé durant ma thèse. Enfin, merci à mes parents qui ont toujours tout fait pour que je puisse choisir ce que je voulais faire dans la vie.

Table des matières

Résumé	i
Abstract	iii
Remerciements	v
Table des matières	vi
Liste des tableaux	ix
Liste des figures	xii
Introduction	1
Lèvre et parole	2
Visiophonie et animation d'avatar	4
Maquillage virtuel	8
Biométrie	9
Historique	10
Objectifs	11
Chapitre 1. État de l'art de l'analyse labiale	15
1.1 Introduction	16
1.2 Les espaces couleur pour l'analyse labiale	16
1.2.1 Espaces couleur classiques	17
1.2.2 Composantes Chromatiques développées pour la segmentation des lèvres	21
1.3 Gradients adaptés à la modélisation du contour des lèvres	23
1.3.1 Gradients basés sur la luminance	24
1.3.2 Gradients hybrides	25
1.3.3 Gradients calculés à partir d'une carte de probabilité	27
1.3.4 Gradients Calculés sur des images binaires	27
1.4 Segmentation des lèvres : Approches Région	27
1.4.1 Approches déterministes	28
1.4.2 Méthodes basées sur la classification	29
1.4.3 Modèles Statistiques de forme et d'apparence	37
1.5 Segmentation des lèvres : Approches Contour	43
1.5.1 Contours actifs	44

1.5.2	Modèles paramétriques	53
1.6	Evaluation des algorithmes de segmentation de la bouche.....	68
1.6.1	Bases d'images de visage	68
1.6.2	Evaluation des performances	70
1.7	Bilan.....	73
Chapitre 2. Segmentation région-contour de la bouche		75
2.1	Introduction.....	76
2.2	Segmentation labiale : Hypothèses de travail	76
2.3	Grandeurs colorimétriques pour la modélisation des lèvres	77
2.4	Gradients Multi-échelle pour la modélisation des contours de la bouche	83
2.5	Segmentation région-contour des lèvres.....	90
2.5.1	Méthodologie	90
2.5.2	Localisation de la zone du visage	91
2.5.3	Localisation et segmentation des lèvres.....	99
2.6	Bilan.....	107
Chapitre 3. Détection de l'état de la bouche.....		111
3.1	Introduction.....	112
3.2	Modélisation du système visuel humain pour le traitement d'image statique....	113
3.2.1	Rétine	114
3.2.2	Cortex V1.....	126
3.3	Identification de l'état de la bouche.....	131
3.3.1	Base d'images de bouche.....	131
3.3.2	Analyse en Composantes Principales (<i>ACP</i>).....	134
3.3.3	Classification supervisée des images de bouche.....	136
3.4	Résultats expérimentaux	138
3.5	Bilan.....	144
Chapitre 4. Etude de la modalité infrarouge pour la segmentation des lèvres		147
4.1	Introduction.....	148
4.2	Rappels sur le rayonnement thermique.....	149
4.3	Analyse du domaine infrarouge et choix d'un système d'acquisition adapté à l'analyse labiale	152

4.3.1	Spectre infrarouge	152
4.3.2	Capteurs en imagerie infrarouge	153
4.3.3	Choix d'un système d'acquisition infrarouge pour l'analyse labiale	156
4.4	Création d'une base d'image visible/infrarouge	159
4.5	Etude du contraste peau/lèvre dans la modalité infrarouge.....	163
4.5.1	Mise en registre des images de bouche	163
4.5.2	Segmentation manuelle des contours internes et externes des lèvres	164
4.5.3	Contraste peau/lèvre dans la modalité infrarouge	165
4.6	Bilan	169
Chapitre 5.	Segmentation des contours externes et internes de la bouche.....	171
5.1	Introduction	172
5.2	Segmentation du contour externe des lèvres	172
5.2.1	Méthodologie	174
5.2.2	Optimisation des contours externe supérieur et externe inférieur.....	175
5.2.3	Recherche des commissures.....	181
5.2.4	Modélisation finale du contour externe de la bouche	186
5.3	Segmentation du contour intérieur de la bouche	187
5.3.1	Modélisation du contour interne pour les bouches ouvertes	187
5.3.2	Modélisation du contour interne des lèvres pour le cas des bouches fermées	197
5.4	Résultats expérimentaux	198
5.4.1	Evaluation des performances basée sur le calcul d'une aire relative	200
5.4.2	Calcul des descripteurs de Fourier pour l'évaluation de la performance de la segmentation.....	203
5.4.3	Cas limites	215
5.5	Bilan	218
Conclusion et perspectives	219	
Travail réalisé	220	
Perspectives	222	
Bibliographie	225	

Liste des tableaux

Table 1.1 : Variances intraclasses et interclasses pour les pixels de peau et de lèvres dans l'espace <i>RGB</i>	18
Table 1.2 : Variances interclasses et intraclasses des distributions des pixels de peau et des lèvres pour les composantes <i>Cb</i> et <i>Cr</i>	20
Table 1.3 : Variances interclasses et intraclasses des distributions des pixels de peau et de lèvres pour les composantes <i>H</i> , <i>H</i> et \hat{U}	21
Table 1.4 : Erreurs de détection sur les points clés (Eveno, 2004).....	71
Table 2.1 : Variances intraclasses, interclasses et V_{intra}/V_{inter} pour la luminance.	79
Table 2.2 : Variance intraclasses, variance interclasses et V_{intra}/V_{inter} pour les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , <i>Hdecorr</i> et \hat{U}_{decorr}	81
Table 2.3 : Variance intraclasses, Variance interclasses et V_{intra}/V_{inter} ainsi que le paramètre w_{max} maximum des mélanges avec $M=[2,3,4,5]$ gaussiennes.	96
Table 3.1 : Résultats de classification lorsque les images de bouche sont traitées par l'ensemble de la chaîne modèle de rétine et modèle de cortex V1 lorsque tous les sujets sont inclus dans l'entraînement du réseau <i>SVM</i>	138
Table 3.2 : Résultats de classification des images de bouche pour le test du « leave-one-out ».	139
Table 3.3 : Résultats de classifications lorsque le modèle de rétine est remplacé par une simple correction de luminance (moyenne nulle et variance unitaire) pour l'apprentissage global.	140
Table 3.4 : Résultats de classifications lorsque le modèle de rétine est remplacé par une simple correction de luminance (moyenne nulle et variance unitaire) pour le test du « leave-one-out ».	141
Table 3.5 : Résultats de classification lorsque le modèle de cortex V1 est remplacé par la méthode <i>LBP</i>	143
Table 3.6 : Résultats de classification lorsque lorsque le modèle de cortex V1 est remplacé par la méthode <i>LBP</i> pour le test du « leave-one-out ».....	143
Table 3.7 : Taux de classification par catégorie d'images de bouche pour la base AR.	144
Table 4.1 : Capteurs les plus utilisés en imagerie infrarouge.	154

Table 4.2 : Variance intraclasses, Variance interclasses et V_{intra}/V_{inter} pour la modalité infrarouge IR et les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , $Hdecorr$ et \hat{U}_{decorr}	165
Table 5.1 : Variance intraclasses, variance interclasses et V_{intra}/V_{inter} pour les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , $Hdecorr$ et \hat{U}_{decorr} pour le cas de la séparation lèvres/intérieur de la bouche.....	189
Table 5.2 : Evaluation de la performance de la segmentation du contour externe par le critère σ_{ext}	201
Table 5.3 : Evaluation de la performance de la segmentation du contour interne par le critère σ_{int}	202
Table 5.4 : Evaluation de la performance de la segmentation du contour externe par comparaison des descripteurs de Fourier pour les images des séries 1, 2, 3.....	206
Table 5.5 : Evaluation de la performance de la segmentation du contour externe par comparaison des descripteurs de Fourier pour les images de la série 4.....	206
Table 5.6 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour les images des séries 1, 2, 3.....	207
Table 5.7 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour les images de la série 4.....	207
Table 5.8 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour l'image de la figure 5.33.	208
Table 5.9 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour l'image de la figure 5.34.	208

Liste des figures

Figure 0.1 : Taux de compréhension de logatomes en fonction du rapport signal/bruit pour 3 cas de figure différents : le son seul, le son seul et les lèvres, et le son et le visage dans son ensemble (Le Goff 1995).....	2
Figure 0.2 : Définition de paramètres géométriques pour les applications de lecture labiale.	4
Figure 0.3 : Illustration des dispositifs du projet TEMPOVALSE.....	5
Figure 0.4 : Points de contrôle du modèle de visage dans le standard MPEG-4	6
Figure 0.5 : Schéma des fonctionnalités du projet TELMA	7
Figure 0.6 : Exemple d'animation d'un modèle facial de type MPEG-4 (Wu, 2002).....	8
Figure 0.7 : Exemple d'avatar (Kuo, 2005)	8
Figure 0.8 : Exemple de détection de contour	9
Figure 0.9 : Exemple d'application d'un maquillage virtuel.....	9
Figure 0.10 : Modèle de visage pour la reconnaissance d'expressions faciales : A gauche, les 5 distances calculées sur le modèle de visage, à droite 2 exemples de reconnaissance d'expression (Hammal, 2007)	10
Figure 1.1 : Tracés des histogrammes correspondant aux distributions des pixels des lèvres (en rouge) et de la peau (en bleu) sur les composantes R , G et B	17
Figure 1.2 : Exemple d'image de bouche, a) image de bouche d'entrée, b) Canal R , c) canal G , d) canal B , e) Luminance Y , f) canal Cb , g) canal Cr	19
Figure 1.3 : Tracés des histogrammes des distributions de nos ensembles de pixels de peau et des lèvres dans Cb et Cr . On donne les histogrammes des distributions correspondant aux pixels des lèvres (en rouge) et de peau (en bleu).	20
Figure 1.4 : Exemples d'image de teinte, a) teinte H , b) teinte \mathbf{H} , c) teinte \hat{U}	21
Figure 1.5 : Tracés des distributions des classes des pixels de peau et des lèvres, a) Tracés dans la composante H , b) Tracés dans \mathbf{H} , c) Tracés dans \hat{U} . On donne les histogrammes des distributions correspondant aux pixels des lèvres (en rouge) et de peau (en bleu).....	22

Figure 1.6 : Représentation du gradient vertical d'une image de bouche. a) Luminance, b) Représentation du gradient vertical dont on ne garde que les valeurs positives, c) Représentation du gradient vertical dont on ne garde que les valeurs négatives.....	24
Figure 1.7 : Images des gradients R_{top} et R_{bottom} , a) image de R/G , b) R_{top} , c) R_{bottom}	25
Figure 1.8 : Images des intensités des gradients G_1 et G_2 , a) G_1 , b) G_2	27
Figure 1.9 : Base d'apprentissage : Lèvres segmentées manuellement	29
Figure 1.10 : Exemple de segmentation des lèvres (Gacon, 2005). De gauche à droite nous présentons, l'image d'entrée, l'image après classification et l'image après réduction du bruit par des opérations morphologiques.....	30
Figure 1.11 : Modélisation des classes de pixels de visage, de lèvre et de fond (Patterson, 2002), a) Estimation des distributions des pixels de visage, de lèvre, et de fond dans l'espace RGB , b) Exemple de segmentation labiale.....	31
Figure 1.12 Exemples de segmentation labiale (Liévin, 2004).....	34
Figure 1.13 : Exemple de segmentation des lèvres (Liew, 2003). Du coin haut-gauche au coin bas-droit, image d'entrée, partition floue de la classe lèvre, partition après des opérations de morphologie mathématique, partition après application des contraintes de symétrie, partition après filtrage gaussien et enfin masque binaire des lèvres après seuillage.....	35
Figure 1.14 : Exemple de modèle de forme de bouche. A gauche, on donne l'ensemble des points clés utilisés pour échantillonner la forme de la bouche. A droite, on donne les 4 principaux modes de variations de la forme de la bouche.	37
Figure 1.15 : Exemple de modèle d'apparence de bouche. A gauche, on donne la grille d'échantillonnage de la texture de la bouche. A droite, on donne les 6 premiers modes de variation du modèle d'apparence.....	40
Figure 1.16 : Schéma de principe de l'algorithme proposé par (Gacon, 2005)	41
Figure 1.17 : Exemples de contours segmentés par l'algorithme présenté par Gacon (Gacon, 2005).....	42
Figure 1.18 : a) Accumulation verticale des pixels sombres, b) Projection horizontale de l'intensité du gradient de la luminance (Delmas 1999).....	47
Figure 1.19 : Les 3 forces externes F_p , F_a et F_r proposées par (Schinchi, 1998).....	51
Figure 1.20 : Exemples d'images de bouche (Martinez, 1998)	54

Figure 1.21 : Modèle de contour interne à 2 paraboles, a) Modèle de contour interne, sur les exemples b) et c), les paraboles sont jointes aux commissures externes de la bouche, sur les exemples d) et e) les paraboles sont jointes aux commissures internes de la bouche.....	55
Figure 1.22 : Modèle de contour interne pour une bouche fermée, a) Modèle de contour interne à une parabole, b) et c) exemples de convergence du modèle.....	56
Figure 1.23 : Modèles de contours internes (Stillittano, 2008), a) Modèles paramétriques pour le contour interne des lèvres (Stillittano, 2008), b) et c) exemples de convergence pour des bouches fermées, d) et e) exemples de convergence du contour pour des bouches ouvertes.....	56
Figure 1.24 : Modèle de contour externe, a) Modèle paramétrique composé de 3 quadriques (Yuille, 1992), b) et c) exemples de convergence du modèle.....	57
Figure 1.25 : Modèle de contour externe à 2 paraboles, a) Modèle paramétrique de bouche à 2 paraboles, b) et c) Exemples de convergence	58
Figure 1.26 : Modèle de contour externe à 3 paraboles, a) Modèle paramétrique à 3 paraboles, b) et c) Exemples de convergence.....	58
Figure 1.27 : Modèle paramétrique proposé par (Eveno, 2004).....	59
Figure 1.28 : Modèles paramétriques basés sur des splines, a) Modèle proposé par Vogt (Vogt, 1996), b) Modèle proposé par Malciu (Malciu, 2000).	59
Figure 1.29 : Détection de l'état de la bouche (Pantic, 2001), a) Rectangles placés dans la zone de bouche, b) profil <i>d</i> de bouche fermée, c) profil <i>d</i> de bouche ouverte (Pantic, 2001)	61
Figure 1.30 : Initialisation du modèle paramétrique (Zhiming, 2002), a) Projections horizontales et verticales de la composante chromatique, b) boîte englobant la bouche.	62
Figure 1.31 : Jumping snake (Eveno, 2004), a) Snake initial, b) Tracés des snakes pour les nouvelles positions du germe, c) Recherche du point bas du contour externe inférieur.	63
Figure 1.32 : Contraintes internes sur le modèle paramétrique (Malciu, 2000), a) Contraintes élastiques, b) contraintes de symétrie.....	64

Figure 1.33 : Exemples d'évaluations subjectives (Kuo, 2005), De droite à gauche, parfait, bon, moyen et mauvais.....	70
Figure 1.34 : Points clés Q_i utilisés pour l'évaluation (Eveno, 2004).....	71
Figure 1.35 : Exemples de vérités-terrain (Stillitano, 2008). A gauche, on donne un exemple de sourire et de cri provenant de la base AR, à droite on donne des exemples d'images acquises avec un casque porté par le sujet et filmant la bouche.....	71
Figure 2.1 : Exemples d'images de bouche utilisées dans notre étude	77
Figure 2.2 : Histogrammes de luminance des ensembles de pixels de la peau et des lèvres.	78
Figure 2.3 : Tracés des histogrammes des distributions des pixels de la peau (en rouge) et des pixels des lèvres (en bleu) pour les grandeurs chromatiques obtenues à partir des grandeurs R_{decorr} , G_{decorr} , B_{decorr} , a) R_{decorr} , b) G_{decorr} , c) B_{decorr} , d) Cb_{decorr} , e) Cr_{decorr} , f) H_{decorr} , g) \mathbf{H}_{decorr} et h) \hat{U}_{decorr}	82
Figure 2.4 : Exemple d'une image de bouche dans RGB , a) image sans traitement, b) image après décorrélation et normalisation.	83
Figure 2.5 : Exemple de calcul du gradient de \hat{U} pour une image de bouche : a) Image de bouche, b) Teinte \hat{U} , c) Intensité du gradient de \hat{U}	84
Figure 2.6 : Calcul du gradient de \hat{U} , pour σ ($t=\sigma^2$) allant de 1 à 10.....	85
Figure 2.7 : Intensité des gradients $\nabla \mathbf{L}(\mathbf{x}, \mathbf{y}, \mathbf{t})$ de \hat{U} le long de la coupe centrale de l'image de bouche pour des valeurs de σ allant de 1 à 10.	86
Figure 2.8 : Intensité des gradients $\nabla \mathbf{normL}(\mathbf{x}, \mathbf{y}, \mathbf{t})$ de \hat{U} le long de la coupe centrale de l'image de bouche pour des valeurs de σ allant de 1 à 10.....	88
Figure 2.9 : Schéma bloc résumant les étapes de la segmentation des lèvres.....	91
Figure 2.10 : Tracés des mélanges de gaussiennes pour $M=[2,3,4]$: a) Image d'entrée, b) Image de teinte \hat{U} , c) Mélange pour $M=2$, d) Mélange $M=3$, e) Mélange pour $M=4$, en rouge, on donne l'histogramme de l'image dans \hat{U} , en vert, on donne les tracés des distributions gaussiennes issues des mélanges.....	97
Figure 2.11 : Résultat de la détection de la zone du visage pour $M=[2,3,4]$	98
Figure 2.12 : Exemple de détection des pixels du visage.....	98
Figure 2.13 : Exemples de masques candidats des lèvres pour une image de bouche, a) Image d'entrée, b) teinte \hat{U} , c) masque du visage, d) masque candidat des lèvres pour	

un seuil trop élevé, e) masque candidat des lèvres pour une valeur de seuil pertinente, les zones en rouge ne sont pas strictement incluses dans le masque du visage, les zones en blanc sont strictement incluses dans le visage.	101
Figure 2.14 : Tracé du plus petit polygone convexe englobant le masque candidat des lèvres de la figure 2.13-e.	102
Figure 2.15 : Tracé de $CP(se)$ avec $N_{echelle} = (3)^2$ pour l'image de bouche de la figure 2.13-a, a) tracé de $CP(se)$, b) masque candidat des lèvres pour le seuil maximisant $CP(se)$	103
Figure 2.16 : Recadrage de la zone de recherche de la bouche sur la zone du visage, a) masque candidat des lèvres ainsi que le tracé des limites haute et basse, b) masque du visage, c) nouvelle zone de recherche.	103
Figure 2.17 : Aire entre le contour convexe de deux masques des lèvres candidats, a) masque candidat à $\varepsilon-1$, b) masque candidat à ε , c) aire entre les 2 masques.	104
Figure 2.18 : Représentation de l'intensité du gradient γ -normalisé $\nabla normL(x, y, t)$, pour $t=1$ sur la ligne verticale passant au milieu de la zone du visage	105
Figure 2.19 : Évolution de $S(t=\sigma^2)$, a) tracé de $S(t=\sigma^2)$, b) images de l'intensité des gradients $\nabla normL(x, y, t)$ pour $1 < \sigma < 6$	106
Figure 2.20 : Exemples de segmentations des lèvres.....	107
Figure 2.21 : Exemples de segmentations des lèvres.....	108
Figure 2.22 : Exemples de segmentations des lèvres pour des sujets ayant la peau noire.	108
Figure 2.23 : Exemples de segmentations erronées.....	109
Figure 3.1 : Schéma du système visuel humain (Benoit, 2007).	113
Figure 3.2 : Schéma de l'œil humain (http://fr.wikipedia.org/wiki/Oeil_humain).....	114
Figure 3.3 : Organisation de la rétine (Benoit, 2007)	115
Figure 3.4 : Réponse d'un photorécepteur en fonction de la lumière reçue dans son voisinage (Kolb, 1996). En trait plein sont tracées les dynamiques d'un photorécepteur en fonction de la luminance locale. En pointillé est tracée la moyenne de la dynamique.....	116
Figure 3.5 : Tracés des réponses d'un photorécepteur pour différentes valeurs de R_0	117
Figure 3.6 : Exemple de compression adaptative des photorécepteurs sur une image de bouche.....	118

Figure 3.7 : Modélisation d'une triade synaptique (Hérault, 2001)	119
Figure 3.8 : Modèle électrique de la <i>PLE</i> (Beaudot, 1994).....	120
Figure 3.9 : Fonction de transfert $G_{PLE}(fs,ft)$	122
Figure 3.10 : Effet des différents réseaux de cellules de la <i>PLE</i> , a) image après compression adaptative, b) filtrage passe-bas des photorécepteurs, c) filtrage passe-bas des cellules horizontales, d) sortie des cellules bipolaires <i>ON</i> , e) sortie des cellules bipolaires <i>OFF</i> , f) sortie <i>ON-OFF</i>	122
Figure 3.11 : Effet de blanchiment spectral de la <i>PLE</i>	123
Figure 3.12 : Signal visuel sur la voie Parvocellulaire en sortie de la rétine, a) image d'entrée, b) sortie <i>ON-OFF</i> sans compression adaptative des cellules ganglionnaire P, c) signal visuel <i>ON-OFF</i> transmis par la voie Parvocellulaire.	124
Figure 3.13 : Modélisation de la rétine pour l'analyse d'images statiques	125
Figure 3.14 : Exemple de calcul d'un spectre Log-polaire. Le spectre d'amplitude de l'information visuelle, issue du modèle de rétine, est échantillonnée par une rosace de filtres <i>Glop</i> pour obtenir le spectre Log-polaire.....	127
Figure 3.15 : Banc de filtres Glop en 1 dimension, à gauche nous avons tracé les profils des filtres sur l'axe des fréquences et à droite sur l'axe des log-fréquences. Les amplitudes des filtres ont été normalisées entre 0 et 1 pour améliorer la visualisation de la forme des filtres.	129
Figure 3.16 : Schéma global des traitements rétine+cortex V1	130
Figure 3.17 : Exemples d'images extraits de la base d'images de bouche segmentées à l'aide de la méthode du chapitre 2, sur la première ligne sont représentées des images de bouche ouverte et sur la seconde ligne des images de bouche fermée.....	132
Figure 3.18 : Effet du fenêtrage de Hanning sur la transformée de Fourier de la sortie Parvocellulaire de la rétine pour une image de bouche.....	133
Figure 3.19 : Transformée de Fourier de la fenêtre de Hanning	134
Figure 3.20 : Projection des spectres Log-polaires des images de bouche sur les 2 premières composantes principales données par <i>ACP</i> . Le nuage de points bleus correspond aux bouches fermées, le nuage de points rouges correspond aux bouches ouvertes.	135
Figure 3.21 : Hyperplan optimal dans le cas d'un problème de séparation linéaire à 2 classes. En rouge est tracée la frontière de décision, les échantillons entourés en bleu	

sont les vecteurs de support.	
(http://fr.wikipedia.org/wiki/Machine_a_vecteurs_de_support).	137
Figure 3.22 : Images de bouche identifiées subjectivement comme fermées et reconnues comme des bouches ouvertes par le classifieur.	139
Figure 3.23 : Images de bouche identifiées subjectivement comme ouvertes et reconnues comme des bouches fermées par le classifieur.	140
Figure 3.24 : Voisinages circulaires (Ojala, 2002)	141
Figure 4.1 : Spectre électromagnétique	149
Figure 4.2 : Spectre de rayonnement de Planck pour différentes températures du corps noir. En rouge, nous avons tracé la courbe à $T=310.15\text{K}$ ce qui correspond à la température du corps humain. En jaune, nous avons tracé la courbe du rayonnement pour $T=5777\text{K}$ qui correspond approximativement à la température à la surface du soleil.	151
Figure 4.3 : Transmittance atmosphérique pour le domaine infrarouge avec une atmosphère standard aux États-Unis en 1976, au niveau de la mer, $T=288\text{K}$, taux d'humidité 46%, pression atmosphérique 1013 millibar, et sur un parcours d'un kilomètre horizontal. En abscisse, on donne la longueur d'onde en μm . En ordonnée, on donne la transmittance en %. Les gaz causant l'absorption sont également indiqués (Wikipedia, 2009).	153
Figure 4.4 : Déetectivité D^* en fonction de la longueur d'onde λ pour différents types de capteurs infrarouges (Maldaque, 2001).	155
Figure 4.5 : Contraste thermique dans les bandes <i>MWIR</i> et <i>LWIR</i> , à gauche, nous avons tracé le contraste pour $T_2 - T_1 = 1\text{K}$, à droite, nous avons tracé le contraste pour des $\Delta(\text{K})$ autour de $T_{\text{corps}}=310.15\text{K}$	158
Figure 4.6 : Exemples d'images tirés de la base conjointe visible/infrarouge	161
Figure 4.7 : Schéma du système d'acquisition de la base d'image de visage visible/infrarouge	162
Figure 4.8 : Annotation d'une image de bouche dans la modalité visible et la modalité infrarouge avec 6 paires de points pour la mise en registre, a) Annotation de l'image visible de départ, b) Annotation de l'image infrarouge, c) Image visible transformée.	163

Figure 4.9 : Segmentation manuelle des lèvres pour le cas d'une bouche ouverte. En noir, nous avons tracé le contour externe et en bleu le contour interne.....	164
Figure 4.10 : Histogrammes normalisés des ensembles de pixels de la peau et des lèvres segmentés manuellement pour la modalité infrarouge.....	166
Figure 4.11 : Exemples d'images de bouches ouvertes dans la modalité infrarouge (<i>IR</i>) et dans la modalité visible (\hat{U} et <i>RGB</i>).....	167
Figure 4.12 : Exemples d'images de bouches fermées dans la modalité infrarouge (<i>IR</i>) et dans la modalité visible (\hat{U} et <i>RGB</i>).....	168
Figure 5.1 : Exemples d'échecs des méthodes de modélisation de la bouche par approche contour : a) Exemple d'échec dû à une mauvaise initialisation, à gauche, on donne le contour initial, à droite, on donne le contour final (Delmas, 2000), b) Exemples de convergence d'un snake avec des paramètres différents (Delmas, 2000), c) et d) Exemples de description du contour de la bouche avec un modèle paramétrique trop simple, les croix représentent le contour réel et les lignes blanches représentent le modèle à 2 paraboles après optimisation.....	173
Figure 5.2 : Schéma bloc des étapes de modélisation du contour externe de la bouche à partir du masque binaire des lèvres.....	174
Figure 5.3 : Exemple de contour externe supérieur initial. En rouge, nous avons tracé la partie supérieure du polygone convexe entourant le masque des lèvres que nous considérons comme le contour initial C_{sup} . Les croix jaunes correspondent aux N_{pt} points $KP_{sup}=[Ksup_1, \dots, Ksup_{Npt}]$ de contrôle.....	176
Figure 5.4 : Déformation de la courbe modélisant le contour C_{sup} avec $deg_{sup}=4$ sous l'action du déplacement d'un point de contrôle symbolisé par une croix bleue. On donne la somme FL des flux des gradients γ -normalisés $\mathbf{VnormL}(x, y, t)$ à travers la courbe ainsi que le déplacement Δ par rapport à la position verticale d'origine du point de contrôle.....	178
Figure 5.5 : Exemples de courbes C_{sup} obtenues avec notre procédure d'optimisation pour des valeurs croissantes de deg_{sup} pour une image de bouche ouverte.....	179
Figure 5.6 : Exemple de contour externe inférieur initial. En rouge, nous avons tracé la partie inférieure du polygone convexe que nous considérons comme le contour initial.	

Les croix jaunes correspondent aux N_{pt} points $KP_{inf} = [Kin_1, \dots, Kin_{N_{pt}}]$ de contrôle.	180
Figure 5.7 : Exemple de contour externe inférieur C_{inf} obtenu après optimisation par notre méthode. En rouge nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.	180
Figure 5.8 : Exemples d'optimisation des contours externe supérieur et externe inférieur pour des images de bouche.	181
Figure 5.9 : Zoom sur les zones des commissures de la bouche, a) zone de la commissure gauche, b) image de bouche, c) zone de la commissure droite	181
Figure 5.10 : Tracés des lignes Lg_{min} et Ld_{min} pour une image de bouche, a) Image d'entrée, b) Zoom sur la zone de la commissure gauche, c) zoom sur la zone de la commissure droite.	183
Figure 5.11 : Tracés des déformations des courbes $\{C'_{sup}, C'_{inf}\}$ lorsque l'indice m augmente pour un point PLg_{min} de Lg_{min} particulier. Les tracés en trait plein correspondent aux courbes $\{C'_{sup}, C'_{inf}\}$ calculées par la méthode des moindres carrés à partir des points de contrôle marqués par des croix.	185
Figure 5.12 : Tracés des couples $\{C'_{sup}, C'_{inf}\}$ pour différents points de Lg_{min} et Ld_{min} . En trait plein blanc, on donne les couples maximisant FL . En pointillés, on donne les tracés des couples candidats, et en vert les tracés de Lg_{min} et Ld_{min} , a) cas de la commissure gauche, b) cas de la commissure droite.	186
Figure 5.13 : Exemple de modélisation du contour externe de la bouche, a) avant l'étape d'optimisation finale, b) après optimisation finale.	186
Figure 5.14 : Tracés des histogrammes des distributions des ensembles de pixels de la zone interne de la bouche (en rouge) et des lèvres (en bleu), a) R_{decorr} , b) G_{decorr} , c) B_{decorr} , d) Cb_{decorr} , e) Cr_{decorr} , f) H_{decorr} , g) $Hdecorr$ et h) \hat{U}_{decorr} .	188
Figure 5.15 : Image de bouche ouverte pour différentes grandeurs chromatiques, a) image RGB , b) R_{decorr} , c) Cr_{decorr} , d) \hat{U}_{decorr} .	189
Figure 5.16 : Images des sommes des intensités des gradients γ -normalisés pour les composantes R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} pour les échelles $t=\sigma^2$ avec $\sigma=\{1,2,3\}$, a) sommes des intensités pour R_{decorr} , c) somme des intensités pour Cr_{decorr} , d) somme des intensités pour \hat{U}_{decorr} .	190

Figure 5.17 : Représentation de la partition Q_{bouche} pour $M_{bouche} = 3$ et de l'image des blobs étiquetés $E(M_{bouche})$, a) Q_{bouche} , b) $E(M_{bouche})$	192
Figure 5.18 : Elimination successive des blobs situés dans la partie supérieure de la bouche et tracés des parties supérieures des polygones convexes entourant les blobs restants.	192
Figure 5.19 : Représentation de E_{haut}	193
Figure 5.20 : Elimination successive des blobs situés dans la partie inférieure de la bouche et tracés des parties inférieure des polygones convexes entourant les blobs restants.	193
Figure 5.21 : Représentation de E_{bas}	193
Figure 5.22 : Masque binaire de la zone interne la bouche.	194
Figure 5.23 : Segmentation de la région interne d'une bouche ouverte pour M_{bouche} croissant.	194
Figure 5.24 : Exemple de contour interne supérieur obtenu après optimisation. En rouge, nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.	195
Figure 5.25 : Exemple de contour interne inférieur obtenu après optimisation. En rouge, nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.	196
Figure 5.26 : Exemple de recherche des commissures internes pour une bouche ouverte, a) Recherche de la commissure interne gauche, b) recherche de la commissure interne droite.	196
Figure 5.27 : Exemple de contour interne final pour une bouche ouverte.	197
Figure 5.28 : Exemple de tracé de L_{min} pour une bouche fermée.	197
Figure 5.29 : Exemple de contour interne final pour une bouche fermée.	198
Figure 5.30 : Exemples d'images tirées des bases d'images.	199
Figure 5.31 : Exemples de vérités-terrain.	200
Figure 5.32 : Exemple de tracé de l'aire A_e entre le contour externe d'une bouche provenant de la vérité-terrain et le contour externe donné par nos algorithmes, on donne également σ_{ext}	200

Figure 5.33 : Tracés des contours internes donnés par la vérité-terrain et par notre algorithme. Le tracé rouge correspond à la segmentation automatique et le tracé blanc correspond à la vérité-terrain. On donne également σ_{int}	202
Figure 5.34 : Tracés des contours internes donnés par la vérité-terrain et par notre algorithme. Le tracé rouge correspond à la segmentation automatique et le tracé blanc correspond à la vérité-terrain. On donne également σ_{int}	203
Figure 5.35 : Exemple de tracé de spectres d'amplitude normalisés des descripteurs de Fourier, pour $\{k=-N_p/2+1, \dots, -1, 2, \dots, N_p/2\}$, du contour interne de la vérité terrain et du contour obtenu par nos algorithmes de segmentation, a) Tracés de la vérité-terrain (blanc) et du résultat de la segmentation (rouge), b) Tracés des spectres d'amplitude normalisés pour la vérité-terrain (croix vertes) et pour le contour obtenu par segmentation (rond rouge).....	205
Figure 5.36 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 1	209
Figure 5.37 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 2	210
Figure 5.38 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 3	211
Figure 5.39 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 4	212
Figure 5.40 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 4 avec des ouvertures extrêmes de la bouche.....	213
Figure 5.41 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de bouche fermée.....	214
Figure 5.42 : Exemples de segmentations erronées du contour externe	215
Figure 5.43 : Exemples de segmentation correcte avec des sujets ayant la peau noire.	216
Figure 5.44 : Exemple de segmentation erronée avec un sujet ayant la peau noire, a) image d'entrée, b) teinte \hat{U} , c) Contour externe segmenté.	216
Figure 5.45 exemples de segmentations erronées du contour interne supérieur dues à la présence des gencives	217
Figure 5.46 : Exemple de segmentation erronée du contour interne inférieure.....	217

Introduction

L'analyse faciale est un domaine très actif de la vision numérique. En effet le développement du marché des nouvelles technologies, comme la téléphonie mobile et la photo numérique, conjugué à l'augmentation du taux d'équipement en ordinateurs particuliers a dynamisé le domaine de la vision notamment pour tout ce qui touche à l'interaction avec les terminaux. Le développement de l'analyse faciale profite également de la croissance du marché de la sécurité depuis les attentats du 11 septembre 2001 aux États-Unis. En effet la zone du visage est un vecteur d'information très important. Les émotions, le langage, l'identité sont autant d'informations portées par le visage. L'accroissement des capacités de calcul a permis, ces dernières années, de développer des méthodes de plus en plus complexes d'analyse faciale. La segmentation du visage en plusieurs régions (peau, cheveux), la modélisation des indices visuels du visage (yeux, bouche, rides,...) et la dynamique de ces caractéristiques sont les traitements classiquement utilisés pour étudier les visages.

Les applications dérivées de l'étude de ces informations sont très nombreuses. Le grand public a pu découvrir depuis quelques années des appareils photos numériques capables de détecter les visages dans les images et d'adapter les réglages des prises de vue en conséquence. Récemment, des fabricants ont proposé des appareils incluant des modules de détection d'émotions basiques comme le sourire. L'analyse des caractéristiques du visage est également largement répandue pour des applications en biométrie. Les yeux, la bouche, la forme du visage sont autant de caractéristiques particulières qui peuvent permettre d'identifier un individu.

La zone du visage particulièrement importante est la bouche, car elle intervient dans la plupart des algorithmes d'analyse faciale et notamment parce qu'elle est liée à la communication et aux émotions. La modélisation de la bouche, en particulier la segmentation des contours des lèvres joue un rôle important dans un grand nombre d'applications.

Lèvre et parole

Les applications les plus directes de la segmentation des contours des lèvres sont liées à la reconnaissance de la parole. En effet, la perception humaine de la parole est fortement influencée par l'information visuelle. Le phénomène le plus courant qui met en évidence cet aspect bimodal de la parole est la capacité de certaines personnes à lire sur les lèvres sans entendre les paroles prononcées.

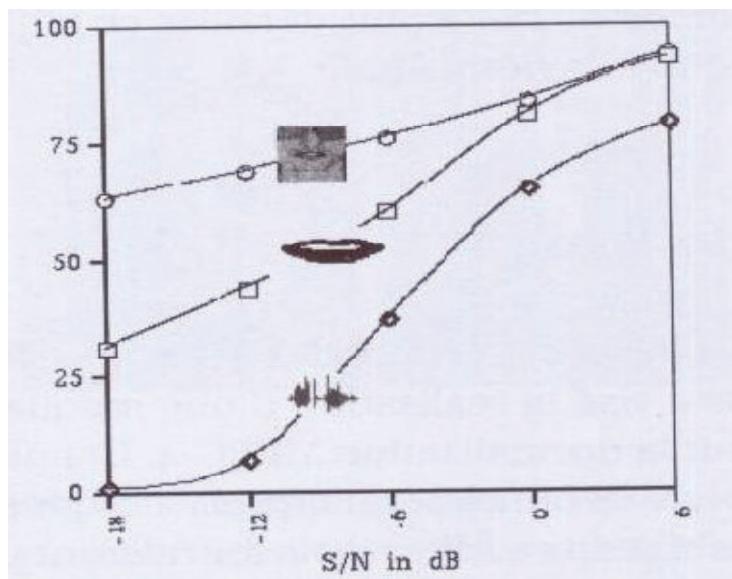


Figure 0.1 : Taux de compréhension de logatomes en fonction du rapport signal/bruit pour 3 cas de figure différents : le son seul, le son seul et les lèvres, et le son et le visage dans son ensemble (Le Goff 1995)

De nombreuses études ont clairement mis en évidence l'aspect bimodal de l'information dans la perception de la parole. Dans (McGurk, 1976) des sujets ont été soumis à des tests dans lesquels étaient présentés des stimuli auditifs accompagnés de stimuli visuels différents ce qui a conduit à la perception de sons différents de ceux présentés. Un formalisme a été créé pour pouvoir étudier la perception de la parole. Du point de vue auditif, la parole peut être décomposée en phonèmes. Il existe 48 phonèmes pour l'anglais (Rabiner, 1993). L'unité élémentaire d'information visuelle est appelée visème, il existe 9 familles de visèmes. D'un individu à un autre, un même phonème peut correspondre à des sons différents, on parle alors d'allophones. L'information visuelle, le visème, peut permettre de percevoir correctement un phonème, là où l'information sonore aurait été

insuffisante. L'information visuelle de la bouche est particulièrement importante dans le cas d'environnements bruités. Dans (MacLeod, 1987) et (Summerfield, 1989) les auteurs ont estimé l'apport de l'information visuelle à un gain de 11dB sur le rapport signal/bruit dans un environnement où l'information sonore est bruitée. Dans (Le Goff, 1995) l'auteur étudie les scores de compréhension dans 3 cas de figure différents : le son seul, le son et les lèvres seules, le son et le visage dans son ensemble (figure 0.1). On remarque sur la figure 0.1 que même, dans le cas d'un signal sonore faiblement bruité, la compréhension est améliorée de 20 % par l'ajout de la modalité visuelle.

De nombreux systèmes de reconnaissance automatique de la parole ont exploité cet aspect bimodal de la perception de la parole. La plupart des algorithmes proposent d'utiliser des indices visuels, particulièrement les contours des lèvres, pour améliorer le taux de reconnaissance. Dans (Potamianos, 2004), l'auteur propose une étude des méthodes classiques de reconnaissance audio-visuelle de la parole. L'information pertinente dans le contexte de la reconnaissance de la parole réside dans la forme et les mouvements de la bouche. Les contours des lèvres doivent être extraits avec une grande précision. Dans (Sugahara, 2000) la segmentation des lèvres est utilisée pour construire un vecteur de caractéristiques contenant 24 paramètres. Douze de ces paramètres correspondent aux distances entre des points régulièrement répartis sur le contour externe par rapport au centre de la bouche. Les douze autres paramètres correspondent aux déviations de ces distances par rapport à l'image précédente. La reconnaissance de la parole est alors basée sur ce vecteur. Shinchi utilise les mêmes 12 points mais analyse les aires des triangles formés par ces points et le centre de la bouche (Shinchi, 1998). Les caractéristiques couramment calculées pour caractériser la forme de la bouche pour la lecture labiale sont basées sur les dimensions de la bouche. Dans (Chan, 1998), (Barnard, 2002), ce sont la largeur et la hauteur du contour qui sont utilisées. Seguier ajoute à ces 2 paramètres leurs dérivées temporelles ainsi que le pourcentage de pixels clairs et le pourcentage de pixels sombres dans la zone de la bouche (Seguier, 2003). Un ensemble plus important de paramètres géométriques, tirés des contours internes et externes de la bouche, est utilisé dans (Zhang, 2002). La figure 0.2 présente les paramètres géométriques classiques extraits des contours de la bouche.

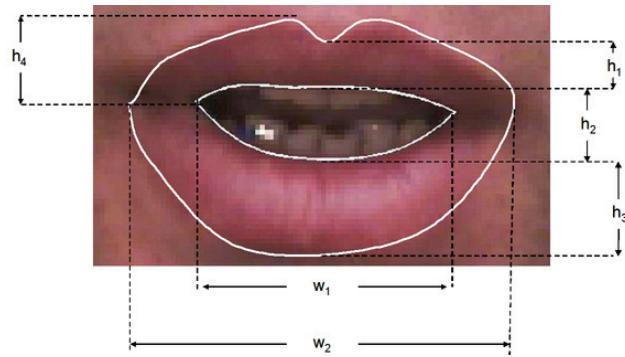


Figure 0.2 : Définition de paramètres géométriques pour les applications de lecture labiale.

Dans (Rosenblum, 1996; Summerfield, 1989; Summerfield, 1992) il est également mis en avant que l'information de texture (présence des dents, de la langue durant la parole) est importante. Les méthodes développées dans (Chan, 2001; Gacon, 2005; Zhang, 2002) fournissent cette information en plus du contour. Dans les travaux de Chan (Chan, 2001) et Zhang (Zhang, 2002) les auteurs montrent que 6 paramètres géométriques combinés avec l'information sur la présence des dents ou de la langue permettent d'obtenir des taux de reconnaissance supérieurs au cas où seuls les paramètres géométriques sont utilisés. Aleksic défend également cette idée (Aleksic, 2002).

Visiophonie et animation d'avatar

Depuis quelques années, nous assistons aux développements de la visiophonie, notamment grâce aux développements d'internet qui a rendu ce type de communication accessible au grand public. Bien que les débits offerts, même aux particuliers, soient toujours plus importants, la qualité des communications, en particulier les communications vidéo, reste très moyenne. En effet, les applications de visiophonie exigent des bandes passantes très importantes même avec un flux vidéo compressé. Des projets ont donc été proposés pour coder les formes et les mouvements du visage, et en particulier ceux de la bouche, pour réduire la quantité de données transitant par le réseau. Ces informations sont par la suite utilisées pour animer un avatar virtuel. Ce principe de compression a été implémenté dans le projet RNRT TEMPOVALSE (Bailly, 2003) auquel ont participé les départements DPC (Département Parole et Cognition) et le département DIS (Département Image et des Signal) du Gipsa-lab (figure 0.3).



Figure 0.3 : Illustration des dispositifs du projet TEMPOVALSE

L’objectif de ce projet était de créer un terminal capturant les mouvements du visage, en particulier, de la bouche, pour animer un avatar virtuel basé sur la norme MPEG-4 (MPEG, 1997). Les paramètres utilisés sont :

- les FDP (Facial Definition Parameters) qui correspondent aux points de contrôle du visage (figure 0.4).
- Les FAP (Facial Animation Parameters) qui codent les informations de mouvement des FDP.
- Les FIT (Facial Interpolation Tables) qui codent la manière d’interpoler les FAP.

Dans le projet TEMPOVALSE les FDP sont déterminés sur la première image de la séquence et le dispositif doit analyser les mouvements du visage pour déterminer les FAP du locuteur qui sont transmis à un autre terminal. Une segmentation du contour des lèvres est nécessaire pour modéliser les mouvements de la bouche.

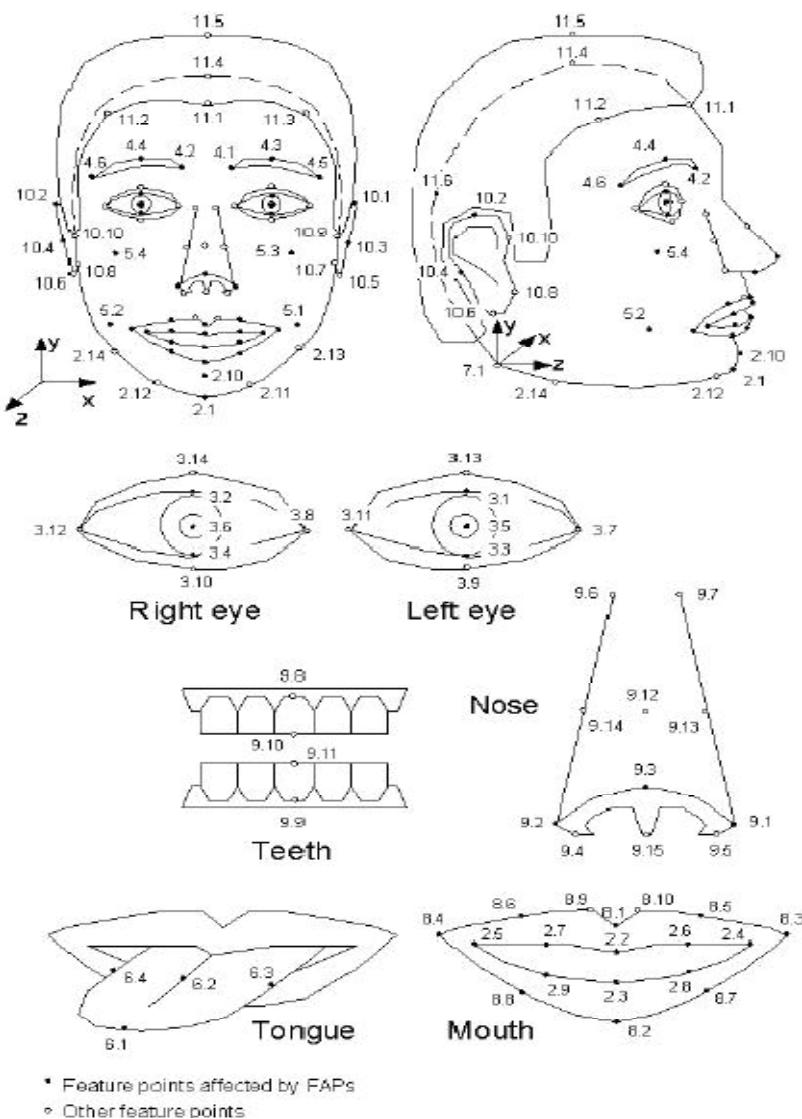


Figure 0.4 : Points de contrôle du modèle de visage dans le standard MPEG-4

L’analyse labiale intervient également dans le projet TELMA (TELéphonie à l’usage des Malentendants) initié par le RNTS (Réseau National des Technologies pour la Santé) auquel a participé le GIPSA-lab (Beautemps, 2007). Le but du projet TELMA était d’utiliser le réseau téléphonique pour permettre aux personnes malentendantes de communiquer à distance. Pour permettre la communication entre malentendantes, le projet TELMA repose sur l’utilisation du LPC en français (Langage Parlé Complété). Le LPC est un code qui associe les postures de la main et des doigts à la parole et qui permet la communication entre bien-entendants et malentendants et de malentendants à

malentendants. Nous savons que pour des individus bien-entendants, l'information visuelle est importante notamment en environnement bruité car elle permet de lever l'ambiguïté qui peut exister sur les allophones. Il existe également des sosies labiaux pour lesquels des phonèmes différents produisent des visèmes analogues. Le LPC permet de lever les ambiguïtés sur les formes de la bouche en complétant l'information grâce à la position de la main et des doigts. Le fonctionnement du projet TELMA est résumé par la figure 0.5. Comme pour le projet TEMPOVALSE, une segmentation des contours de la bouche intervient pour modéliser la bouche et ses mouvements. Les mouvements des doigts de la main sont aussi modélisés. Ces informations sont transmises sur le réseau et servent à animer un clone virtuel sur le terminal du correspondant ainsi qu'à faire la synthèse d'un signal de parole. Dans le cadre d'une communication entre individus bien-entendants, une segmentation labiale permet de débruiter le signal sonore transmis au correspondant (Rivet, 2006).

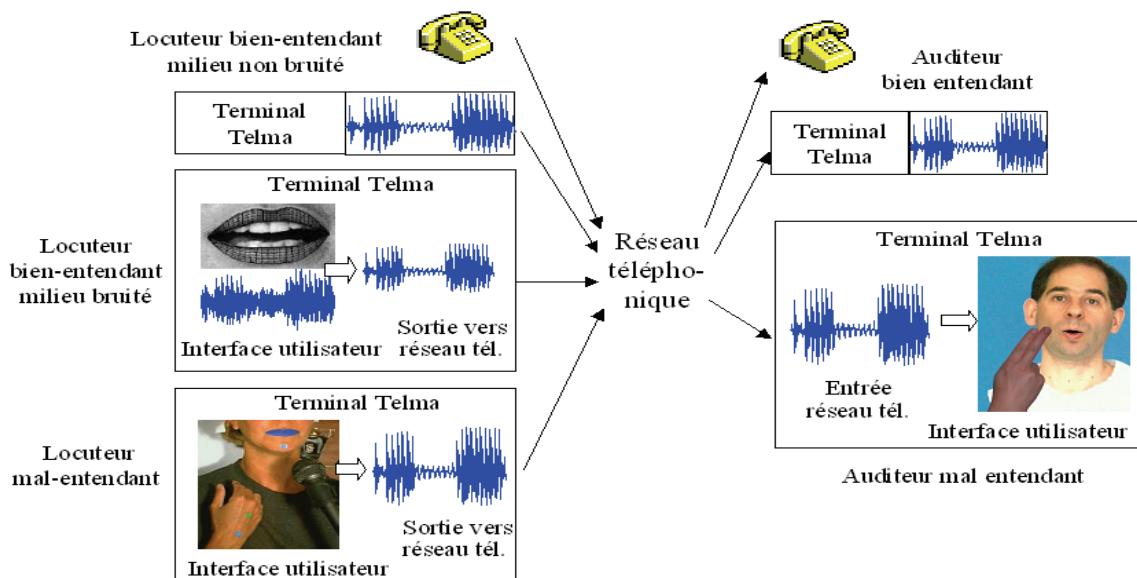


Figure 0.5 : Schéma des fonctionnalités du projet TELMA

La segmentation des lèvres dans le but d'animer des clones virtuels a également été étudiée dans (Beaumesnil, 2006; Kuo, 2005; Yin, 2002; Wu, 2002). Dans (Wu, 2002 ; 2004) les

auteurs utilisent un modèle de visage analogue à celui de la norme MPEG-4 dont les déformations de la bouche sont guidées par un algorithme de segmentation des lèvres (figure 0.6).



Figure 0.6 : Exemple d'animation d'un modèle facial de type MPEG-4 (Wu, 2002)

Dans (Kuo, 2005) le contour des lèvres segmenté permet d'animer un modèle en 3 dimensions du visage du sujet (figure 0.7). Yin utilise aussi la segmentation des lèvres pour animer un modèle en 3 dimensions de visage (Yin, 2002).



Figure 0.7 : Exemple d'avatar (Kuo, 2005)

Maquillage virtuel

L'analyse labiale intervient également dans des applications de réalité augmentée. La société Vesalis développe un système de maquillage virtuel en temps réel. Le traitement logiciel met en œuvre des opérations de détection des contours du visage, des yeux, de la bouche et des sourcils (figure 0.8). L'interface permet ensuite à l'utilisateur d'appliquer un maquillage virtuel sur les différentes zones du visage. Dans ce contexte, l'extraction des contours de la bouche demande une grande précision de manière à appliquer le maquillage sur les zones concernées uniquement (figure 0.9).

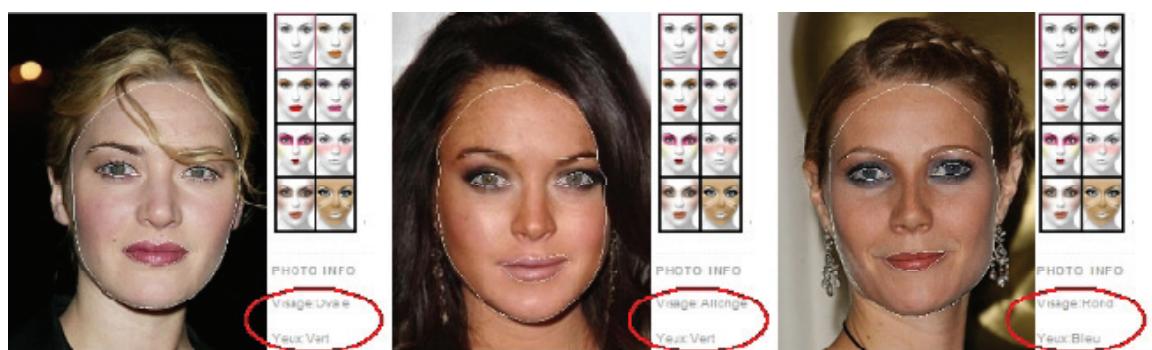


Figure 0.8 : Exemple de détection de contour



Figure 0.9 : Exemple d'application d'un maquillage virtuel.

Biométrie

La segmentation labiale intervient aussi dans le domaine de la biométrie et de la reconnaissance d'expressions. Dans (Brand, 2001; Chibelushi, 1997; Hsu, 2002; Luettin, 1997; Wark, 1998) les auteurs étudient le potentiel biométrique des lèvres. L'extraction du

contour des lèvres est employée pour déterminer des paramètres géométriques sur les bouches. Ces paramètres sont ensuite utilisés pour l'identification de personnes.

Dans (Hammal, 2006; Hammal, 2007) 5 paramètres géométriques sont calculés sur un modèle de visage incluant les lèvres, les yeux et les sourcils (figure 0.10). Ces 5 mesures géométriques sont utilisées dans un système de classification d'expression faciale.

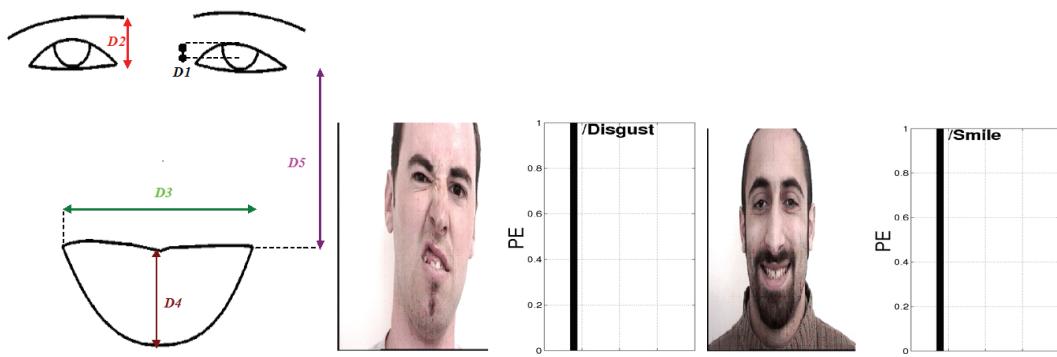


Figure 0.10 : Modèle de visage pour la reconnaissance d'expressions faciales : A gauche, les 5 distances calculées sur le modèle de visage, à droite 2 exemples de reconnaissance d'expression (Hammal, 2007).

De la même manière, des paramètres géométriques sont calculés à partir des contours des lèvres dans (Seyedarabi, 2006) pour classifier les images des lèvres en unité d'action (Action Units (Ekman 1978)). Enfin dans (Yokogawa, 2007), le contour des lèvres est utilisé pour planifier des opérations de chirurgie plastique.

Historique

L'analyse labiale est un thème de recherche actif au GIPSA-lab depuis le début des années 90 avec les travaux de (Lallouache, 1991) à l'ICP, aujourd'hui devenu le DPC (Département Parole et Cognition). Dans ces travaux, un maquillage bleu était utilisé pour segmenter les lèvres dans des séquences vidéo. Cette activité s'est ensuite poursuivie au DIS (Département Image et Signal) avec les travaux de (Lievin, 2000) qui a développé une méthode basée sur un modèle de champs de Markov aléatoires utilisant des grandeurs colorimétriques et le mouvement pour segmenter les contours internes et externes des lèvres. Delmas (Delmas, 2000) a quant à lui développé un algorithme utilisant des contours

actifs pour extraire les contours de la bouche alors que Eveno (Eveno, 2003) a abordé le problème de la segmentation et du suivi des lèvres par des modèles paramétriques déformables. Dans (Gacon, 2005), les recherches ont porté sur les modèles de formes actifs (ASM) et les modèles d'apparences actifs (AAM) pour modéliser la zone de la bouche et extraire les contours externes et internes des lèvres ainsi que les dents dans un contexte multi-locuteurs.

Objectifs

Les exemples d'applications précédents montrent que l'analyse labiale est un thème largement abordé à l'heure actuelle et que l'extraction des contours de la bouche fait partie des traitements mis en œuvre dans de nombreuses applications d'analyse faciale. À l'heure actuelle, malgré le nombre important de méthodes de segmentation des lèvres qui existent, le problème reste ouvert. Certaines applications nécessitent en effet une grande précision dans la modélisation du contour et une grande robustesse par rapport aux changements de conditions de l'environnement et de sujets. De tels algorithmes sont très dépendants de la précision de la segmentation. Comme tout problème de segmentation, et en particulier lié à l'humain, l'extraction du contour des lèvres est un problème complexe. Les algorithmes doivent pouvoir gérer des lèvres qui subissent des déformations importantes et très rapides lors de la prononciation d'une phrase. Les algorithmes doivent également gérer des conditions de l'environnement qui peuvent être très différentes et des orientations des sujets variables. L'objectif visé dans notre étude est de proposer un ensemble de méthodes permettant de modéliser précisément la zone de la bouche avec la meilleure robustesse possible. Par robustesse, nous entendons obtenir une méthode fiable qui ne nécessite pas de réglage de paramètres et qui réalise une segmentation fidèle des contours externes et internes des lèvres. Nous verrons qu'il existe à l'heure actuelle 2 grandes familles de méthodes pour segmenter les lèvres : Les méthodes dites « régions » et les méthodes dites « contours ». Nous présenterons dans ce rapport notre approche combinée « région-contour » pour effectuer une segmentation des contours des lèvres.

Ce rapport présente les travaux réalisés dans le cadre d'une thèse effectuée en cotutelle entre l'Institut Polytechnique de Grenoble et l'Université Laval à Québec. Les travaux ont impliqué les laboratoires universitaire GIPSA-lab à Grenoble et le Laboratoire de Vision et

Systèmes Numériques de l'Université Laval (LVSN). Le premier chapitre sera consacré à l'état de l'art des techniques d'analyse labiale. Tout d'abord, nous étudierons les principales grandeurs colorimétriques qui sont utilisées pour le problème de l'extraction des contours des lèvres. Par la suite, nous nous intéresserons aux deux grandes familles de méthodes de segmentation des lèvres : les approches « régions » et les approches « contours ». Nous nous attacherons à présenter les méthodes principales appartenant à ces 2 familles.

Le second chapitre introduira une approche combinée région-contour dans le but d'obtenir une segmentation multi-locuteur de la bouche sur des images de visage en couleurs. Nous décrirons en premier lieu une technique permettant d'augmenter le contraste entre la peau et les lèvres. Par la suite, un formalisme multi-échelle sera proposé pour améliorer la robustesse de la modélisation des contours de la bouche. Ensuite, nous présenterons notre méthode de localisation et de segmentation de la bouche sur des images de visage.

Le chapitre 3 sera consacré à la détection de l'état de la bouche. Cette étape est nécessaire à la modélisation de la région intra labiale qui présente une grande variabilité de forme et d'apparence. Une approche bio-inspirée basée sur un modèle de rétine et de cortex visuel a été développée pour résoudre ce problème d'identification. La première section introduira les modèles de rétine et de cortex visuel. La deuxième section du chapitre se concentrera sur la méthode d'identification de l'état de la bouche. La dernière partie du chapitre sera consacrée à l'évaluation de notre méthode d'identification de l'état de la bouche.

Ces dernières années de nombreux auteurs ont proposé d'utiliser la modalité infrarouge en analyse faciale, particulièrement pour la reconnaissance de visage. Le LVSN possédant une solide expertise dans le domaine de la vision infrarouge, il nous a paru pertinent d'étudier l'intérêt de cette modalité pour le problème de la segmentation des lèvres dans le cadre de la cotutelle de thèse. Le chapitre 4 décrit nos travaux sur le potentiel de la modalité infrarouge dans le cadre de la segmentation des lèvres. Notre intérêt pour cette modalité porte sur l'amélioration de la robustesse de la segmentation. Nous étudierons dans un premier temps la bande infrarouge du spectre électromagnétique pour déterminer les bandes de fréquences pertinentes en fonction des contraintes de notre problème et des contraintes imposées par les systèmes d'acquisition. Nous donnerons un descriptif de la base d'images de visages combinée visible/infrarouge que nous avons créée pour mener à bien notre

étude. Enfin, nous présenterons notre étude sur le gain apporté par la modalité infrarouge pour séparer les lèvres et la peau.

Puis, dans le chapitre 5 nous détaillerons la modélisation et la segmentation des contours externes et internes de la bouche. La première partie de ce chapitre sera consacrée à la modélisation du contour externe de la bouche. Nous introduirons un modèle de contour dont la complexité sera automatiquement adaptée en fonction de la bouche traitée. L'extraction des contours internes des lèvres est particulièrement difficile car l'aspect de la zone interne de la bouche peut varier très fortement. Divers éléments (dents, langue, cavité buccale) peuvent être visible ou non. Pour surmonter cette difficulté, nous introduirons une méthode de classification non-supervisée destinée à la sélection des régions internes de la bouche. Notre méthode de modélisation de contour sera par la suite appliquée aux contours internes de la bouche. La dernière partie du chapitre sera consacrée à l'analyse quantitative des performances de notre segmentation. Nous présenterons une méthode d'évaluation originale basée sur des descripteurs de forme pour pouvoir comparer efficacement les contours détectés à des vérités-terrain.

Finalement, nous présenterons nos conclusions sur les algorithmes proposés et des perspectives pour la poursuite de nos travaux sur l'analyse labiale.

Chapitre 1. État de l'art de l'analyse labiale

1.1 Introduction

Dans ce chapitre, nous présentons un état de l'art des méthodes d'analyse labiale. L'étude qui suit est consacrée à la modélisation des lèvres sur des séquences d'images de visage vu de face et dans lesquelles les lèvres sont toujours visibles.

La section 1.2 traitera le problème du choix de l'espace couleur qui a une importance capitale sur la performance de la segmentation des lèvres. La section 1.3 sera centrée sur les gradients proposés pour la modélisation des contours des lèvres. La section 1.4 donnera un aperçu des méthodes de modélisation des lèvres dites « région ». Dans la section 1.5, nous nous concentrerons sur les approches dites contours. Enfin, la section 1.6 présentera les méthodes classiques d'évaluation des algorithmes de segmentation labiale.

1.2 Les espaces couleur pour l'analyse labiale

Le choix d'une grandeur colorimétrique pertinente est déterminant pour la performance de tout algorithme de segmentation. Pour le problème de la segmentation des lèvres, on cherchera à travailler avec une grandeur qui sépare au mieux la classe des pixels des lèvres de la classe des pixels de peau. D'un point de vue statistique, cela revient à chercher un espace dans lequel les variances intraclasses sont minimales et les variances interclasses maximales.

Dans la sous-section 1.2.1 nous reviendrons sur les espaces couleur couramment utilisés en traitement d'image (espaces *RGB*, *HSV*, *YCbCr*). Dans la sous-section 1.2.2 nous nous intéresserons à l'étude d'espaces spécifiques qui ont été proposés afin d'améliorer le contraste entre la peau et les lèvres. Nous étudierons le pouvoir discriminant de chaque grandeur colorimétrique en analysant les distributions des pixels de peau et de lèvres. Une base d'images spécifique a été construite pour cette étude. Cette base est constituée de 150 images de 20 sujets différents. Pour toutes ces images, les pixels de peau et des lèvres ont été manuellement extraits. Ces images ont été acquises avec la même caméra et avec des conditions d'illumination équivalentes. Pour comparer les dynamiques des classes de pixels de peau et des lèvres dans les différents espaces, les composantes chromatiques ont été d'abord normalisées entre [0,1] et les variances interclasses V_{inter} et intraclasses V_{intra} ainsi que le ratio V_{intra}/V_{inter} ont été calculés. Le ratio V_{intra}/V_{inter} sera d'autant plus petit que la

variance intraclasses sera petite et que la variance interclasses sera grande. Soit un ensemble de N échantillons répartis en k classes, les variances intraclasses V_{intra} et interclasses V_{inter} sont calculées de la manière suivante :

$$\begin{aligned} V_{intra} &= \sum_{h=1}^k \frac{n_h}{N} V^h \\ V_{inter} &= \sum_{h=1}^k \frac{n_h}{N} (\bar{X}^h - \bar{X})^2 \end{aligned} \quad (1)$$

avec n_h le nombre d'échantillons dans la classe h , V^h la variance de la classe h , \bar{X}^h la moyenne de la classe h et \bar{X} la moyenne de l'ensemble des échantillons.

1.2.1 Espaces couleur classiques

1.2.1.1 Espace *RGB*

On se propose d'étudier la pertinence de l'espace *RGB* pour la segmentation des lèvres sur notre base de test. Dans la Figure 1.1, on donne les histogrammes des pixels de peau et des lèvres pour les 3 composantes couleur R , G et B . Les variances intraclasses, interclasses et V_{intra}/V_{inter} sont donnés dans la table 1.1.

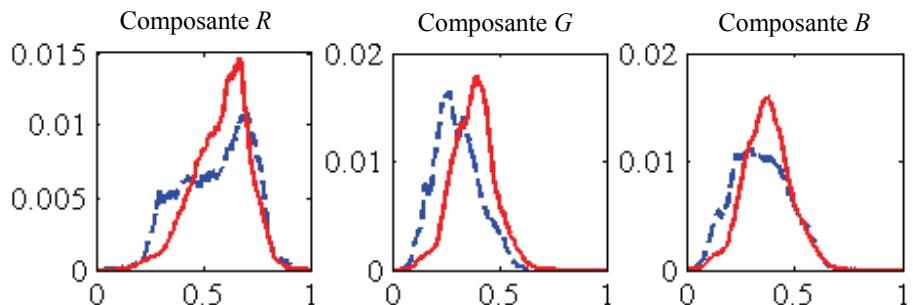


Figure 1.1 : Tracés des histogrammes correspondant aux distributions des pixels des lèvres (en rouge) et de la peau (en bleu) sur les composantes R , G et B .

On rappelle que les histogrammes de la figure 1.1 représentent les distributions des pixels des lèvres et de peau pour un ensemble d'images provenant de 20 sujets différents. Les tracés des histogrammes ainsi que les résultats de la table 1.1 mettent donc en valeur les tendances des ensembles de pixels de peau et des lèvres sur plusieurs sujets. On remarque

immédiatement qu'il y a un fort recouvrement entre les distributions de couleur des pixels de peau et de lèvres pour chacune des composantes couleur. Les résultats des calculs de variances intraclasses et interclasses montrent également un fort recouvrement entre les 2 classes de pixels. Dans la figure 1.2 sont présentées, à titre d'exemple, les composantes *RGB* d'une des images de la base de test. On remarque sur l'image d'entrée en couleurs que les lèvres semblent globalement plus rouge que la peau. Sur les images des composantes *RGB* présentées à la figure 1.2, il semble que la séparation est meilleure dans la composante *G*. Dans la composante *R* on constate visuellement que la séparation est beaucoup moins nette et que les niveaux des pixels de la lèvre supérieure et de la lèvre inférieure ne sont pas homogènes, ce qui indique une sensibilité aux variations de luminance. La table 1.1 confirme cette impression visuelle, on constate que V_{intra}/V_{inter} est le plus faible pour *G* et le plus grand pour *R*. Globalement les variances interclasses sont faibles par rapport aux variances intraclasses. D'après les tracés des histogrammes et les résultats de la table 1.1, il est évident que l'espace *RGB* n'est pas adapté à la segmentation des lèvres. Il y a un très fort recouvrement entre les distributions des classes de pixels de lèvres et de peau. Ces résultats indiquent une très faible stabilité des propriétés des classes de pixels de peau et des lèvres d'une image à l'autre. Il sera donc très difficile d'obtenir un traitement robuste en travaillant avec cet espace. A noter que dans cet espace, l'information de luminance est corrélée avec les composantes chromatiques. Le fort recouvrement entre les distributions des classes de pixels des lèvres et de peau est lié à cet aspect. Nous comparerons ces résultats dans la suite à ceux des autres espaces.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
Composante <i>R</i>	0.0185	1.0245e-004	180.1447
Composante <i>G</i>	0.0097	0.0010	9.6215
Composante <i>B</i>	0.0123	1.6431e-004	74.8480

Table 1.1 : Variances intraclasses et interclasses pour les pixels de peau et de lèvres dans l'espace *RGB*.

1.2.1.2 Espace *YCbCr*

L'espace *YCbCr* est dérivé de l'espace *RGB* (Ford, 1998). Cet espace a été créé à l'origine pour séparer l'information de luminance des composantes chromatiques en proposant une transformation linéaire et bijective découpant la luminance de la chrominance. Cette

transformation aboutit au calcul de 3 composantes : Y qui correspond à l'information de luminance et $[Cb,Cr]$ qui sont les composantes chromatiques.

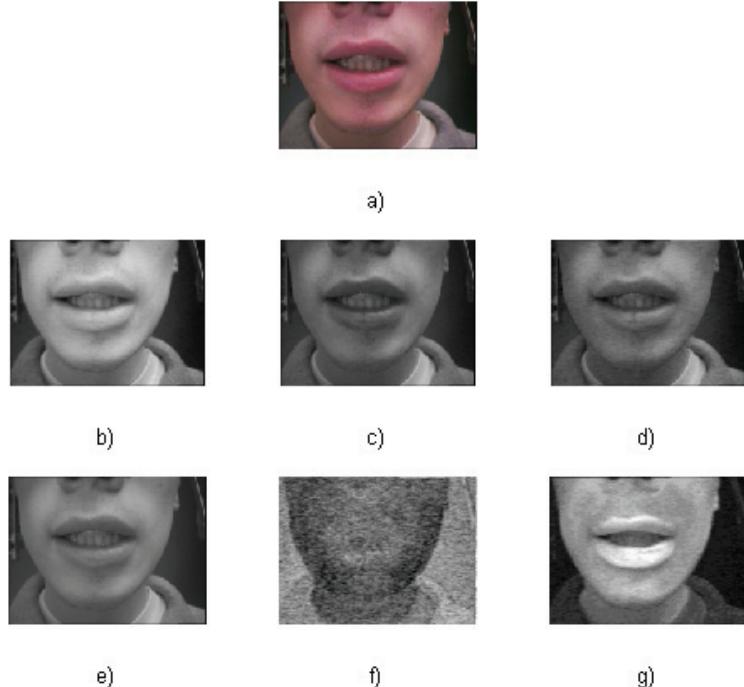


Figure 1.2 : Exemple d'image de bouche, a) image de bouche d'entrée, b) Canal R , c) canal G , d) canal B , e) Luminance Y , f) canal Cb , g) canal Cr .

La figure 1.2 présente également les composantes Y , Cb et Cr d'une image de bouche. Visuellement la composante Cb (figure 1.2-f) semble contenir très peu d'information exploitable pour la segmentation des lèvres. La composante Cr (figure 1.2-g) semble au contraire offrir un contraste plus important entre les lèvres et la peau du visage du sujet et les différentes zones sont relativement homogènes. La figure 1.3 présente les histogrammes des distributions de nos ensembles de pixels de peau et des lèvres dans Cb et Cr . On observe encore un fort recouvrement entre les distributions des pixels à la fois dans Cb et Cr . Les résultats de la table 1.2 montrent une légère amélioration de la séparation peau/lèvre par rapport aux composantes RGB , mais les variances intraclasses restent supérieures aux variances interclasses. Globalement pour l'espace $YCbCr$ les rapports V_{intra}/V_{inter} sont plus faibles que pour RGB mais il subsiste néanmoins toujours un très fort recouvrement entre les distributions.

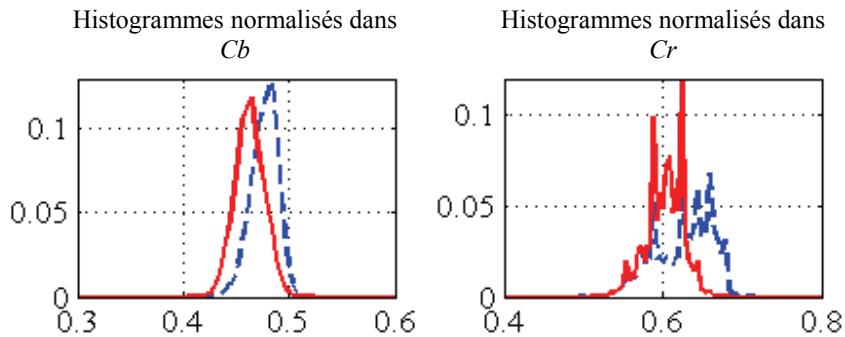


Figure 1.3 : Tracés des histogrammes des distributions de nos ensembles de pixels de peau et des lèvres dans Cb et Cr . On donne les histogrammes des distributions correspondant aux pixels des lèvres (en rouge) et de peau (en bleu).

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
Composante Cb	$1.8208 \cdot 10^{-4}$	$3.3567 \cdot 10^{-5}$	5.4245
Composante Cr	$7.0541 \cdot 10^{-4}$	$8.5732 \cdot 10^{-5}$	8.2281
Y	0.0128	$3.3172 \cdot 10^{-4}$	38.4922

Table 1.2 : Variances interclasses et intraclasses des distributions des pixels de peau et des lèvres pour les composantes Cb et Cr .

1.2.1.3 Espace HSV , HSI , HSL

Les espaces tels que HSV (pour Hue / teinte, Saturation / saturation, Value / valeur), HSI (pour Hue / teinte, Saturation / saturation, Intensity / intensité) et HSL (pour Hue / teinte, Saturation / saturation, Lightness / luminance) où la chrominance et la luminance sont aussi séparées, ont été également utilisés pour la segmentation des lèvres (Zhang, 2000 ; Coianiz, 1996). Bien que les expressions des transformations vers ces espaces soient différentes, les composantes chromatiques codent des informations similaires. H code l'information de teinte. S code l'information de saturation des couleurs qui correspond à la pureté des couleurs. Les composantes V , I ou L codent, quant à elles, l'information de luminance. Les problèmes de segmentation labiale se basent couramment sur la teinte H .

Dans (Zhang, 2000) les auteurs ont comparé le pouvoir discriminant des espaces RGB , HSV et $YCbCr$ pour la segmentation des lèvres. Après avoir étudié les histogrammes calculés à partir de plusieurs séquences, Zhang et al. ont conclu que la teinte H de l'espace HSV est pertinente pour séparer les pixels de peau et des lèvres (Zhang, 2000). Les auteurs affirment également que H est robuste aux variations de lumière. La figure 1.4-a présente la teinte H

pour une image de bouche. Comme les valeurs de teinte H sont homogènes à des angles et que les valeurs correspondant aux teintes rouges sont proches de 2π , nous avons permuted les valeurs de teinte pour centrer la dynamique sur les niveaux de teinte des lèvres. A l'aide de nos ensembles de pixels de peau et des lèvres nous avons tracé les histogrammes normalisés et recentrés pour la teinte H (figure 1.5-a). On donne par ailleurs la variance intraclasses, la variance interclasses et V_{intra}/V_{inter} dans la table 1.3. Ces résultats montrent une amélioration par rapport aux composantes de $YCbCr$ étudiées précédemment. Il subsiste, néanmoins, toujours un recouvrement important entre les distributions des pixels de peau et des pixels des lèvres.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
Composante H	$4.8031 \cdot 10^{-004}$	$1.3192 \cdot 10^{-004}$	3.6408
Composante \hat{H}	$5.4321 \cdot 10^{-004}$	$3.1940 \cdot 10^{-004}$	1.7007
Composante \hat{U}	0.0035	0.0020	1.7601

Table 1.3 : Variances interclasses et intraclasses des distributions des pixels de peau et de lèvres pour les composantes H , \hat{H} et \hat{U} .

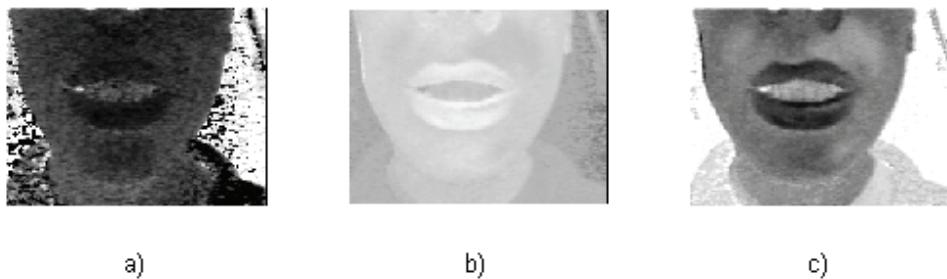


Figure 1.4 : Exemples d'image de teinte, a) teinte H , b) teinte \hat{H} , c) teinte \hat{U}

1.2.2 Composantes Chromatiques développées pour la segmentation des lèvres

Les espaces couleur étudiés précédemment (RGB , $YCbCr$, HSV), s'ils sont utilisés assez largement dans les problèmes d'analyse faciale, ne semblent pas pertinents pour la modélisation des lèvres. Pour améliorer la performance de la segmentation des lèvres, plusieurs grandeurs colorimétriques ont été proposées, en particulier la pseudo-teinte (Poggio, 1998) et la transformation LUX (Liévin, 2004).

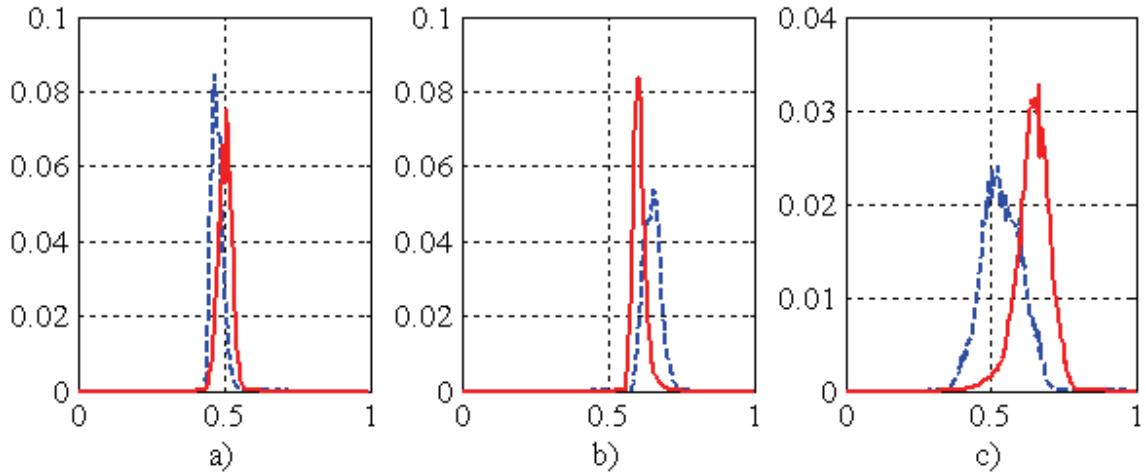


Figure 1.5 : Tracés des distributions des classes des pixels de peau et des lèvres, a) Tracés dans la composante H , b) Tracés dans \hat{H} , c) Tracés dans \hat{U} . On donne les histogrammes des distributions correspondant aux pixels des lèvres (en rouge) et de peau (en bleu).

1.2.2.1 La pseudo-teinte \hat{H}

La pseudo-teinte \hat{H} , proposée par (Poggio, 1998), a été utilisée par Eveno et al. (Eveno, 2003) pour la modélisation des lèvres. Poggio ayant remarqué que les valeurs de R et G étaient plus grandes pour les pixels des lèvres que pour les pixels de peau, la transformation suivante a été proposée, à partir des grandeurs RGB , pour augmenter le contraste entre les lèvres et la peau :

$$\hat{H} = \frac{R}{R+G} \quad (2)$$

Dans la figure 1.5-b sont présentés les histogrammes des pixels de peau et des lèvres. Les variances interclasses et intraclasses ainsi que V_{intra}/V_{inter} sont également fournies dans la table 1.3. On voit immédiatement que V_{intra}/V_{inter} est plus faible que pour H . Le contraste entre les lèvres et la peau est donc plus important qu'avec H . On peut également remarquer sur l'exemple de calcul de \hat{H} donné à la figure 1.4-b que le bruit est beaucoup moins important dans les zones sombres. Cette propriété est intéressante pour les cas où les images sont sombres ou quand la bouche est ouverte, l'intérieur de la bouche étant souvent sombre.

1.2.2.2 La composante \hat{U} de l'espace couleur *LUX*

Liévin et al. ont proposé une autre grandeur chromatique dans le but de maximiser la discrimination entre lèvre et peau (Liévin, 2004):

$$\hat{U} = \begin{cases} 256 \times \frac{G}{R} & \text{if } R > G \\ 255 & \text{otherwise} \end{cases} \quad (R, G, B) \in [0, 256] \times [0, 256] \times [0, 256] \quad (3)$$

Cette transformation est une simplification appliquée au cas de la segmentation labiale de l'espace couleur *LUX* introduit dans (Liévin, 2004). Cette formulation a été inspirée par des considérations biologiques et par le traitement d'image logarithmique (Deng, 1993) dans le but de maximiser le contraste des images pour des problématiques d'analyse faciale. Les variances interclasses, intraclasses et V_{intra}/V_{inter} pour cette teinte sont données à la table 1.3. La figure 1.4-c présente un exemple de calcul de \hat{U} pour une image de bouche et nous avons tracé les histogrammes des distributions des ensembles de pixels de peau et de lèvres à la figure 1.5-c. On voit que V_{intra}/V_{inter} est légèrement plus grand que pour \hat{H} . Sur l'exemple de la figure 1.4-c, on constate une séparation importante entre les pixels de peau et des lèvres. Visuellement, on remarque également sur la figure 1.4-c que le contraste est important entre le visage et les autres parties de l'image (fond, vêtements,...). L'exemple de la figure 1.4-c montre également que les zones sombres sont peu bruitées.

Nous pouvons également mentionner le travail de Chiou qui a utilisé une teinte $Q=R/G$ très similaire à \hat{U} pour la localisation de la bouche sur des images de visage (Chiou, 1997). La différence vient de l'absence de condition sur le rapport G/R .

1.3 Gradients adaptés à la modélisation du contour des lèvres

Lorsque l'on s'attaque au problème de la modélisation du contour des lèvres, on cherchera à renforcer les transitions entre les lèvres et les autres régions du visage. Il sera important de travailler avec un gradient dont l'intensité est forte et constante sur les contours afin d'obtenir une modélisation robuste des lèvres. De nombreuses expressions du gradient ont été proposées pour modéliser les contours des lèvres.

1.3.1 Gradients basés sur la luminance

Les gradients les plus largement utilisés pour l'extraction des contours des lèvres sont dérivés de la luminance (Hennecke, 1994; Radeva, 1995; Pardas, 2001; Delmas, 2002; Seyedarabi, 2006; Werda, 2007). En effet, dans le cas du contour externe des lèvres, la transition peau/lèvre est marquée par une variation d'illumination due à des propriétés de réflexion de la lumière sur la peau différentes de celles des lèvres. La morphologie du sujet peut aussi être une cause de variation de luminance entre la peau et les lèvres. Par exemple, dans le cas d'une source lumineuse située au dessus du sujet, la lèvre supérieure sera plus sombre que la zone de peau. De même, une ombre se formera sous la lèvre inférieure, ce qui accentuera la transition entre la zone de lèvre et la zone de peau. Dans le cas du contour intérieur de la bouche, les gradients basés sur la luminance caractérisent aussi très bien la jonction entre les deux lèvres. Pour le cas d'une bouche fermée, la jonction entre les lèvres est très sombre. Le contraste est donc très fort avec les lèvres. Quand la bouche est ouverte, les dents, caractérisées par une zone claire, ou la cavité buccale, caractérisée par une zone sombre, induisent des changements d'intensité lumineuse importants avec les lèvres et donc des gradients forts au niveau des transitions.

Classiquement, les gradients sont composés de la dérivée horizontale et verticale de la luminance. Dans certaines applications, seule la composante verticale est étudiée, étant donné que les contours de la bouche sont majoritairement horizontaux. De plus suivant le signe du gradient on pourra sélectionner le type de transition désirée, une transition d'une zone claire vers une zone sombre ou inversement.

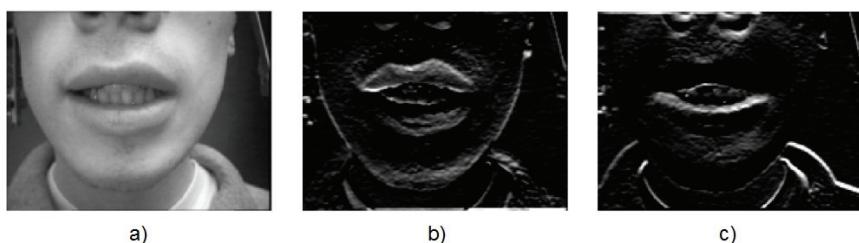


Figure 1.6 : Représentation du gradient vertical d'une image de bouche. a) Luminance, b) Représentation du gradient vertical dont on ne garde que les valeurs positives, c) Représentation du gradient vertical dont on ne garde que les valeurs négatives.

Les gradients basés sur la luminance sont pertinents pour la modélisation des contours internes et externes des lèvres mais, par définition, ils vont être sensibles aux variations d'illumination et des contours non désirés peuvent alors apparaître. Les ombres et les réflexions, sur la peau en particulier, peuvent engendrer des réponses importantes lors du calcul du gradient à partir de la luminance. La figure 1.6-c met en lumière ce problème, la lèvre inférieure projette une ombre qui engendre un gradient élevé sous la lèvre. De plus, Radeva a également noté que la transition entre la lèvre inférieure et la peau est généralement plus douce ce qui engendre des gradients moins forts et qui rend le contour plus difficile à extraire (Radeva, 1995).

En ce qui concerne la zone interne de la bouche, quand celle-ci est ouverte, la présence des dents, de la langue ou simplement de la cavité buccale peut également engendrer des contours parasites. L'algorithme d'extraction des contours des lèvres devra donc être capable de sélectionner les contours désirés. En particulier, nous serons intéressés à extraire la transition entre les lèvres et l'intérieur de la bouche.

1.3.2 Gradients hybrides

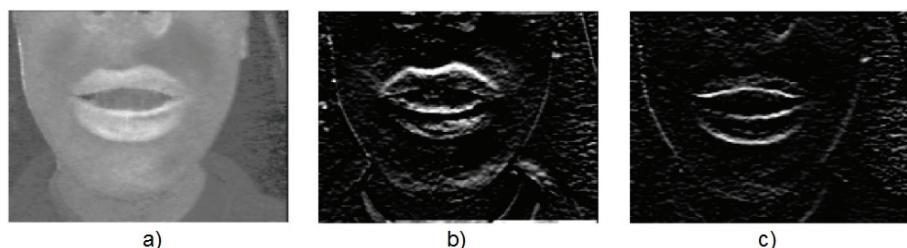


Figure 1.7 : Images des gradients R_{top} et R_{bottom} , a) image de R/G , b) R_{top} , c) R_{bottom} .

D'autres auteurs tels que (Eveno, 2004) ont proposé de combiner différentes informations pour caractériser les contours des lèvres. Eveno et al. ont proposé l'utilisation de gradients spécifiques pour caractériser le contour haut (R_{top}) et le contour bas des lèvres (R_{bottom}) (figure 1.7-b et 1.7-c) :

$$\begin{aligned} R_{top}(x,y) &= \nabla \left(\frac{R}{G}(x,y) - I(x,y) \right) \\ R_{bottom}(x,y) &= \nabla \left(\frac{R}{G}(x,y) \right) \end{aligned} \quad (4)$$

I correspond à la luminance, R et G correspondent aux composantes rouge et verte de l'espace *RGB*. L'hypothèse est que le ratio R/G est plus fort pour les pixels des lèvres (figure 1.7-a) et que la luminance est plus grande pour les pixels de peau que pour les pixels des lèvres.

De la même manière, Beaumesnil et al. utilisent une combinaison de \widehat{U} et de la luminance pour calculer le gradient (Beaumesnil, 2006). Stillittano propose deux expressions du gradient, G_1 et G_2 , respectivement pour le contour intérieur haut et intérieur bas de la bouche (Stillittano, 2008):

$$G_1(x, y) = \nabla(R(x, y) - u(x, y) - \widehat{H}(x, y)) \quad (5)$$

$$G_2(x, y) = \nabla(I(x, y) + u(x, y) + \widehat{H}(x, y))$$

où R correspond à la composante rouge de l'espace *RGB*, I est la luminance, \widehat{H} correspond à la pseudo-teinte et u est une composante provenant de l'espace *CIELuv*. L'espace *CIELuv* est un modèle de représentation des couleurs développé en 1976 par la Commission internationale de l'éclairage (CIE). Une couleur est alors caractérisée à l'aide d'un paramètre d'intensité (luminance L) et de deux paramètres de chrominance (u et v). Ce système est de type perceptif, c'est-à-dire qu'il a été créé pour que les distances entre les couleurs correspondent aux différences perçues par l'œil humain.

Les expressions des gradients G_1 et G_2 sont justifiées par les hypothèses suivantes :

- I et \widehat{H} sont généralement plus grands pour les pixels de lèvres que pour les pixels situés à l'intérieur de la bouche.
- u est plus grand pour les pixels des lèvres que pour les pixels des dents (en effet la valeur de u pour ces pixels est proche de zéro).
- R est plus grand pour les pixels de lèvres que pour les pixels de l'intérieur de la bouche.

La figure 1.8 présente un exemple de calcul de G_1 et de G_2 sur une image de bouche.

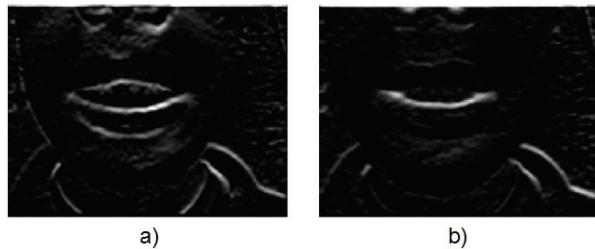


Figure 1.8 : Images des intensités des gradients G_1 et G_2 , a) G_1 , b) G_2

1.3.3 Gradients calculés à partir d'une carte de probabilité

Dans (Vogt, 1996) une carte de probabilité est calculée à partir des composantes H et S de l'espace HSV . Cette carte donne des valeurs élevées pour les pixels des lèvres. Cette carte est ensuite utilisée pour calculer un gradient employé pour l'extraction du contour des lèvres. L'efficacité de cette méthode dépendra alors de la qualité de l'estimation de la probabilité d'appartenance d'un pixel à la classe lèvre.

1.3.4 Gradients Calculés sur des images binaires

Liévin et al. (Liévin, 2004) construisent également une carte de probabilité par une approche utilisant un modèle de champs aléatoires de Markov (*MRF*, Markov Random Field) appliqué sur une combinaison de la teinte \hat{U} et d'un champ de mouvement. Le modèle de champs aléatoires permet de partitionner l'image en plusieurs classes de pixels. Le gradient peut ensuite être calculé sur l'image binaire correspondant aux pixels identifiés comme appartenant aux lèvres. Cette méthode d'extraction du contour a également été utilisée dans (Wark, 1998 ; Yokogawa, 2007). L'avantage de cette méthode est que la modélisation du contour n'est pas parasitée par d'autres contours et que l'optimisation ne se fera que sur des contours identifiés comme étant ceux des lèvres.

1.4 Segmentation des lèvres : Approches Région

Les principales méthodes de segmentation labiale basée sur des approches région peuvent être classées en trois catégories : les méthodes déterministes basées sur la couleur et des techniques de seuillage, les méthodes de classification supervisées et non-supervisées où, par exemple, les pixels sont classifiés en tant que peau et lèvre. Enfin, il existe une dernière

catégorie qui englobe les méthodes dites statistiques basées sur l'apprentissage de modèles de forme et d'apparence de la bouche.

1.4.1 Approches déterministes

Cette catégorie regroupe les méthodes bas niveau pour lesquelles il n'existe pas de connaissance a priori sur les statistiques de l'image ou de modèle, de forme ou colorimétrique, de la zone de la bouche. La segmentation des lèvres résulte alors principalement d'une étape de seuillage sur la luminance ou sur une grandeur colorimétrique particulière. La méthode la plus simple, mais qui est aussi la plus efficace, est basée sur un seuillage couleur (chroma keying). Les lèvres sont maquillées avec une couleur en fort contraste avec la peau, le bleu est par exemple très approprié. Les lèvres sont alors très facilement segmentées en appliquant un seuillage sur H (Lallouache, 1991 ; Chibelushi, 1997 ; Brandt 2001). Ce type de méthode conduit à une segmentation très précise, mais l'utilisation de ce type d'artifice n'est pas envisageable pour les applications courantes. Chiou utilise aussi une étape de seuillage pour segmenter les lèvres. Etant donnée une image centrée sur la zone de la bouche, l'idée est d'obtenir une image binaire des lèvres en effectuant un double seuillage sur la grandeur colorimétrique $Q=R/G$ et sur R (Chiou, 1997). Wark et al. utilisent une approche analogue sur un corpus d'images de visage en couleurs avec des seuils haut et bas sur Q choisis de manière empirique (Wark, 1998). Coianiz et al. (Coianiz, 1996) utilisent un intervalle fixe sur H pour extraire les lèvres. Zhang applique quant à lui des seuils fixes sur H et S pour extraire la zone des lèvres d'une série d'image de visage de sujets différents (Zhang, 2000).

La difficulté principale rencontrée lors de l'utilisation de ces méthodes de bas niveau est la détermination de manière automatique du ou des seuils pour la segmentation des lèvres. Des seuils fixes ne seront pas robustes à des changements des conditions de l'environnement ou des systèmes d'acquisition. Ces méthodes sont donc la plupart du temps utilisées dans le cas d'environnements contrôlés ou lorsque l'application permet le réajustement des seuils par un expert.

1.4.2 Méthodes basées sur la classification

Si l'on fait l'hypothèse que l'image de départ est centrée sur le bas du visage, la segmentation des lèvres peut alors être vue de manière simplifiée comme un problème de classification à deux classes : la classe lèvre et la classe peau. Les techniques de classification (ou de clustering) sont largement utilisées dans les problèmes d'analyse faciale et en particulier pour la segmentation des lèvres. Le problème revient à grouper des pixels dans des ensembles homogènes. Les méthodes classiques employées en analyse faciale font appel à des outils d'analyse statistique (théorie de l'estimation, champs de Markov aléatoires), aux réseaux de neurones, aux machines à vecteurs de support (Support Vector Machine, *SVM*), aux C-moyennes (C-means) et plus récemment aux C-moyennes floues (Fuzzy C-means). Dans la suite, nous présentons des exemples d'application de ces méthodes à la segmentation des lèvres.

1.4.2.1 Méthodes de classification supervisées

Les méthodes de classification supervisées impliquent un modèle de connaissance a priori disponible pour les classes dans lesquelles nous cherchons à placer nos données d'entrée. En général, on utilise les connaissances apprises sur un ensemble d'exemples pour ensuite classifier des données inconnues. La première étape d'un algorithme de classification supervisée pour la segmentation labiale est donc la constitution d'une base d'apprentissage d'images de bouche. La création de la base d'apprentissage est une étape critique. Pour que l'apprentissage soit le plus robuste possible, la base doit couvrir le maximum de cas possibles avec des conditions d'environnement variables.

- Approches statistiques :



Figure 1.9 : Base d'apprentissage : Lèvres segmentées manuellement



Figure 1.10 : Exemple de segmentation des lèvres (Gacon, 2005). De gauche à droite nous présentons, l'image d'entrée, l'image après classification et l'image après réduction du bruit par des opérations morphologiques.

Gacon et al. (Gacon, 2005) modélisent la zone de la bouche par des mélanges de gaussiennes estimées sur une base d'images de bouche manuellement segmentée (Figure 1.9). Tout d'abord, \hat{H} est calculée pour chaque image. Les distributions de couleur de la zone de la bouche sont alors estimées sur l'ensemble des images de la base d'apprentissage par un mélange de gaussiennes pour chaque classe (classe peau et classe lèvre) par un algorithme espérance-maximisation (*EM*). Le nombre optimal de gaussiennes pour chaque classe est donné par le principe de la longueur de description minimale (*MDL*, Minimum Description Length). Dans le cas de Gacon, ce nombre est d'une distribution gaussienne par classe. Une image inconnue est alors traitée de la manière suivante : chaque pixel dans $\hat{H}(x, y)$ est associé à la classe c_i pour laquelle la probabilité d'appartenance $p(\hat{H}(x, y)|c_i)$ est maximale (Figure 10). La probabilité d'appartenance d'un pixel est contrainte par $p(\hat{H}(x, y)|c_i) > p_0$, p_0 est fixé empiriquement. Des opérations morphologiques sont ensuite effectuées pour réduire le bruit sur les masques obtenus (cf. figure 1.10).

Dans (Sadeghi, 2002), les auteurs effectuent également l'estimation de la distribution de couleur de la zone de la bouche par un mélange de gaussiennes afin de classifier les pixels en 2 classes : lèvre et non-lèvre. Pour cela, Sadeghi travaille sur un espace où les composantes R , G , B sont normalisées par rapport à la luminance. Les distributions des pixels lèvre et non-lèvre sont entraînées sur une base d'images de bouche segmentées manuellement pour obtenir une connaissance a priori. Pour une image de bouche inconnue, l'idée est d'estimer le mélange de gaussiennes optimal qui modélise la zone de la bouche en utilisant les modèles de distributions qui ont été entraînés sur une base d'image. Dans cette étude, les auteurs considèrent les composantes chromatiques normalisées pour l'estimation de la mixture de gaussiennes. Pour trouver le nombre optimal de gaussiennes, les auteurs calculent les 2 probabilités suivantes :

$$p_{emp}(x) = \frac{k(W)}{N} \text{ et } p_{pred}(x) = \int_W p(x)dx \quad (6)$$

Où x est un vecteur composé des valeurs (r,g) du pixel considéré, p_{emp} est la probabilité empirique, p_{pred} est la probabilité prédictive par la mixture de gaussiennes courante, W est une fenêtre générée aléatoirement dans l'espace d'observation (r,g) et $k(W)$ est le nombre total de points dans W . Enfin, N est le nombre total de pixels de la zone de la bouche. Le nombre initial de gaussiennes est fixé à 1. Le nombre de gaussiennes est incrémenté de 1 jusqu'à minimisation de l'erreur quadratique entre p_{emp} et p_{pred} sur W . La distribution optimale estimée peut alors aboutir à un grand nombre de gaussiennes, et donc de classes différentes. La distribution initiale entraînée sur une base d'images est alors utilisée pour fusionner les différentes gaussiennes de manière à obtenir un nombre de classes prédéterminé, ici deux classes. Pour cela, les gaussiennes estimées sur l'image inconnue sont considérées comme modélisant les lèvres ou les pixels non-lèvre en fonction de l'erreur entre leurs moyennes et les moyennes des classes données par le modèle supervisé.

Dans (Patterson, 2002), les auteurs utilisent également une base d'images de visage en couleurs segmentées manuellement dans le but d'estimer les distributions des classes de pixels de visage, des lèvres, et du fond par des lois gaussiennes dans l'espace *RGB* (figure 1.11-a). Pour cette étude, les auteurs ont utilisé les images de la base de données *CUAVE* qui inclut des images où le haut du corps, jusqu'aux épaules des sujets, est visible (figure 1.11-b). Une segmentation des lèvres est ensuite réalisée en utilisant une règle de décision de type bayésienne (figure 1.11-b).

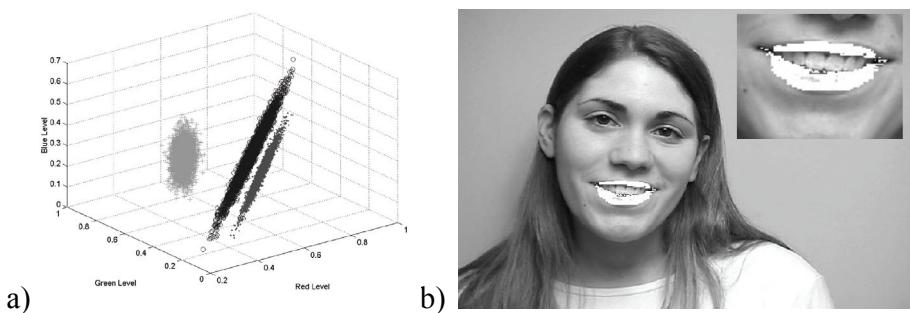


Figure 1.11 : Modélisation des classes de pixels de visage, de lèvre et de fond (Patterson, 2002), a) Estimation des distributions des pixels de visage, de lèvre, et de fond dans l'espace *RGB*, b) Exemple de segmentation labiale.

Au lieu de construire des modèles de distribution de couleur des lèvres ou du visage par apprentissage, d'autres auteurs ont proposé d'utiliser les connaissances apprises sur les classes de pixels des lèvres et de peau pour obtenir des transformations maximisant la séparation peau/lèvres. Chan entraîne un modèle linéaire dans le but de maximiser le contraste entre les lèvres et la peau du visage sur une base d'apprentissage (Chan, 1998). Les paramètres optimaux α , β et γ sont obtenus par optimisation de la composante $C = \alpha R + \beta G + \gamma B$ en maximisant la séparation entre les classes des pixels des lèvres et de peau. Ensuite, une étape de seuillage, avec un seuil fixe, est appliquée pour segmenter les lèvres. Nefian applique une technique similaire dans (Nefian, 2002). Une analyse discriminante linéaire (*LDA*, Linear Discriminant Analysis) est effectuée sur des données provenant d'images segmentées manuellement pour calculer la transformation optimale de *RGB* vers un espace à une dimension dans lequel le contraste entre la peau et les lèvres est maximal. La segmentation est ensuite effectuée par une étape de seuillage de la composante obtenue.

- *Réseaux de neurones :*

Les réseaux de neurones ont également été employés pour résoudre le problème de la classification supervisée des pixels de peau et des lèvres. Les réseaux à propagation vers l'avant comme les perceptrons multicouches ont, en particulier, été utilisés pour la segmentation labiale. Un perceptron multicouche permet d'appliquer des frontières de décisions complexes dans un espace de dimension *ad hoc*. Ce type d'approche est donc intéressant pour modéliser des ensembles de pixels de lèvre et de peau où les pixels sont représentés par des vecteurs. Dans (Wojdel, 2001), les auteurs ont développé un algorithme de segmentation labiale dans lequel une étape d'apprentissage est effectuée en premier lieu. Tout d'abord, le sujet est invité à segmenter manuellement ses lèvres dans le but d'entraîner un perceptron multicouche, qui servira ensuite à classifier les pixels sur le reste de la séquence. Le réseau est composé de trois couches, avec trois neurones sur la première couche, cinq sur la seconde et un neurone en sortie. Les trois entrées du réseau correspondent aux niveaux dans *RGB* des pixels. Le réseau est entraîné pour que la sortie du réseau soit nulle pour les pixels des lèvres et qu'elle retourne 1 pour les pixels du visage.

Daubias et al. emploient aussi des perceptrons multicouches pour classifier les pixels de peau, des lèvres et les pixels de la zone intérieure de la bouche (Daubias, 2002). Afin d'effectuer cette classification, l'architecture générale des réseaux est la suivante : Perceptron multicouche à 3 couches, avec 3 neurones sur la couche de sortie, une sortie pour les pixels des lèvres, 1 sortie pour les pixels de peau et une sortie pour les pixels de la zone interne de la bouche. En entrée du réseau, un pixel est caractérisé par un voisinage $n \times n$ dans *RGB*. Ceci donne un vecteur de dimension $3n^2$ en entrée du réseau. Le réseau est entraîné à partir d'images segmentées manuellement. Les auteurs étudient ensuite la performance de cette architecture pour la classification en fonction du nombre de neurones sur la première couche et sur la couche cachée du réseau.

- *Machine à vecteurs de support ou SVM (Support Vector Machine)*

Les réseaux *SVM* permettent également de placer des frontières de décision entre des ensembles de données. Ils peuvent donc être utilisés pour séparer les pixels des lèvres et de peau. Castañeda et al. utilisent un classifieur de type *SVM* (cf. section 3.3.3) pour détecter des visages et des lèvres dans des images de visage (Castañeda, 2005). Une base de voisinages de taille 10×20 de zones caractéristiques du visage (lèvre, yeux, sourcils, ...) est construite à partir d'images provenant de 13 sujets différents. Le réseau *SVM* est alors entraîné pour classifier les différentes caractéristiques étudiées, dont la bouche. Ensuite, pour une image inconnue, une étape de prétraitement basée sur une segmentation couleur identifie la zone de visage et des zones candidates, caractérisées par des boîtes rectangulaires, pour les différents indices visuels recherchés. Le classifieur *SVM* est alors utilisé pour classifier les zones candidates.

1.4.2.2 Approches non-supervisées

Dans la sous-section précédente, nous avons précisé que la construction d'une base d'apprentissage était une étape critique pour les algorithmes de segmentation supervisée. Yang a étudié le rendu des couleurs obtenu avec différentes caméras dans des conditions contrôlées (Yang, 1996). Les résultats montrent d'importantes variations du rendu de la couleur d'une caméra à l'autre. Ces résultats mettent en lumière une des difficultés rencontrées lors de la création d'une base d'image : la généralisation d'un modèle dans le

cas d'un environnement non contrôlé est très délicate car le modèle doit pouvoir gérer les variations de nombreux paramètres tels que la lumière, l'échelle, la morphologie des sujets ou les changements de système d'acquisition (caméra mono CCD, 3-CCD, CMOS, ...). La segmentation manuelle de centaines, voire de milliers, d'images représente également une difficulté majeure. Pour essayer de contourner cette difficulté, les méthodes de classification non-supervisée sont utilisées pour la segmentation labiale. Ces méthodes ne nécessitent pas d'étape d'apprentissage.

- Approches statistiques

Dans (Liévin, 2004), l'auteur propose un algorithme hiérarchique de segmentation non-supervisée d'un nombre quelconque de classes basé sur un modèle de champs de Markov aléatoires ou *MRF* (Markov Random Fields) utilisé sur des séquences dynamiques d'images de visage. L'intérêt du modèle *MRF* est de combiner l'information de teinte \hat{U} avec un champ de mouvement. L'hypothèse est que le champ de mouvement est important sur la zone des lèvres par rapport au reste du visage lorsque le sujet parle.

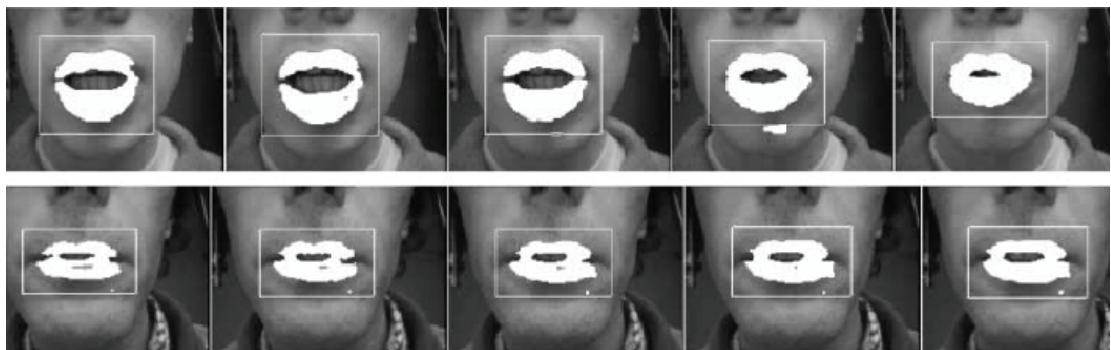


Figure 1.12 Exemples de segmentation labiale (Liévin, 2004).

L'algorithme est initialisé avec un nombre de classes $n=1$. À l'itération n , la première étape est l'estimation de la moyenne μ_n et de l'écart type σ_n du mode principal de l'histogramme des pixels non classifiés dans \hat{U} . On calcule également le champ de mouvement $fd(x,y)=|I_t(x,y)-I_{t-1}(x,y)|$, I_t étant la luminance de l'image courante et I_{t-1} celle de l'image précédente dans la séquence, (x,y) sont les coordonnées dans l'image. À partir de (μ_n, σ_n) et

de $fd(x,y)$, un étiquetage initial $L(x,y) = (n,k)$ est calculé. Les pixels sont initialisés à la classe n courante par seuillage de la teinte $h_n(x,y)$ obtenue par l'expression suivante :

$$h_n(x,y) = \left[256 - \left(\frac{\hat{U}(x,y) - \mu_n}{\sigma_n} \right)^2 \right] \quad (7)$$

avec $|\hat{U}(x,y) - \mu_n| \leq 16\sigma_n$

$k=1$ lorsque le pixel est en mouvement et $k=0$ sinon. L'état de mouvement k est obtenu par seuillage entropique sur fd . Le modèle de champs de Markov aléatoires est ensuite employé pour optimiser l'étiquetage de l'image. L'algorithme est itéré jusqu'à ce qu'il n'y ait plus de pixels à étiqueter. Pour le cas de la segmentation des lèvres, Liévin utilise des images centrées sur la bouche (figure 1.12). N est fixé à 2, la première classe correspond alors à la peau et la seconde aux lèvres avec la contrainte $\mu_2 < \mu_1$. Les pixels de la peau sont considérés fixes. Des exemples de segmentation sont présentés sur la figure 1.12.

- Approches basées sur les K-moyennes et les K-moyennes floues

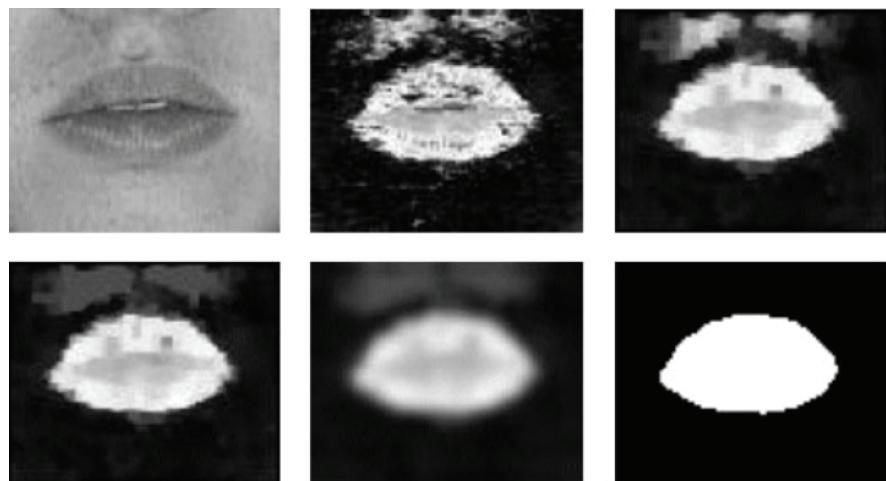


Figure 1.13 : Exemple de segmentation des lèvres (Liew, 2003). Du coin haut-gauche au coin bas-droit, image d'entrée, partition floue de la classe lèvre, partition après des opérations de morphologie mathématique, partition après application des contraintes de symétrie, partition après filtrage gaussien et enfin masque binaire des lèvres après seuillage.

Les algorithmes des K-moyennes (C-means) et des K-moyennes floues (Fuzzy K-means) sont également largement utilisés dans les problèmes de classification non-supervisée. Ces méthodes ont pour but d'effectuer une partition des données d'entrée en N classes. Avec l'algorithme des K-moyennes, cette partition est binaire, une donnée d'entrée appartient à une seule classe. La méthode des K-moyennes floues, contrairement aux K-moyennes, permet d'attribuer des degrés d'appartenance compris entre 0 et 1. Une donnée d'entrée peut donc appartenir à plusieurs classes pourvu que la somme des degrés d'appartenance soit de 1. Dans les deux cas, la partition globale est obtenue par minimisation d'une fonction objective. Dans (Liew, 2003), un algorithme des K-moyennes floues est utilisé pour la segmentation des lèvres. Les images utilisées sont centrées sur la zone de la bouche et le nombre de classes est fixé à 2. Les pixels sont représentés par un vecteur $x_{r,s}$ (r,s étant les coordonnées spatiales) qui comprend les niveaux dans les espaces CIELuv et CIELab. Le critère de performance classique consiste à sommer les normes euclidiennes des vecteurs de données $x_{r,s}$ par rapport aux centres des classes v_i et à pondérer le total par le degré d'appartenance. Dans (Liew, 2003), ce critère a été modifié pour tenir compte du voisinage direct de chaque pixel $x_{r,s}$ dans le calcul de distance par rapport aux centres v_i des classes. Les auteurs considèrent un voisinage 3×3 autour de $x_{r,s}$. La norme $\|x_{r,s} - v_i\|_2$ est aussi pondérée par la somme des normes $\|x_{r+l_1, s+l_2} - x_{r,s}\|_2$, avec $(l_1, l_2) \in [-1, 1]$, voir (Liew, 2003) pour l'expression détaillée du critère de performance. Après la classification, des post-traitements tels que des opérations morphologiques et l'application de contraintes de symétrie sont effectués. Un seuil, fixé à l'avance, est ensuite appliqué sur la partition floue pour segmenter les lèvres (figure 1.13).

Dans (Leung, 2004) et (Wang, 2007), des approches analogues sont utilisées pour classifier des pixels de peau et des lèvres en ajoutant cette fois une contrainte sur la forme de la répartition spatiale des pixels. L'hypothèse est que la forme d'une bouche est elliptique. L'algorithme des K-moyennes a également été utilisé pour la segmentation des lèvres dans (Beaumesnil, 2006) pour classer les pixels d'images de bouche en 3 classes, lèvre, visage et arrière-plan. Pour cette classification, Beaumesnil considère uniquement la composante \hat{U} .

1.4.3 Modèles Statistiques de forme et d'apparence

Les modèles statistiques de forme appartiennent à la catégorie des méthodes supervisées, une base d'apprentissage doit donc être créée. Les modèles statistiques de forme sont entraînés pour décrire les variations de forme et d'apparence d'une zone d'intérêt. L'idée est alors d'optimiser un modèle de forme ou d'apparence pour segmenter la bouche dans une image inconnue. Les premiers travaux utilisant ce type de modèle ont d'abord tenté de segmenter la bouche en utilisant un modèle de forme actif ou *ASM* (Active Shape Model). Des modèles d'apparence actifs, ou *AAM* (Active Appearance Model), ont ensuite été employés pour ajouter l'information de texture à la forme.

1.4.3.1 Modèle de forme actif

Les modèles *ASM* sont dérivés du modèle de distribution de points ou *PDM* (Point Distribution Model) introduit par Cootes (Cootes, 1992 ; 1995). Le modèle *PDM* est entraîné en utilisant un ensemble de points clés échantillonnant la forme de la caractéristique étudiée, donc dans notre cas la bouche (figure 1.14).

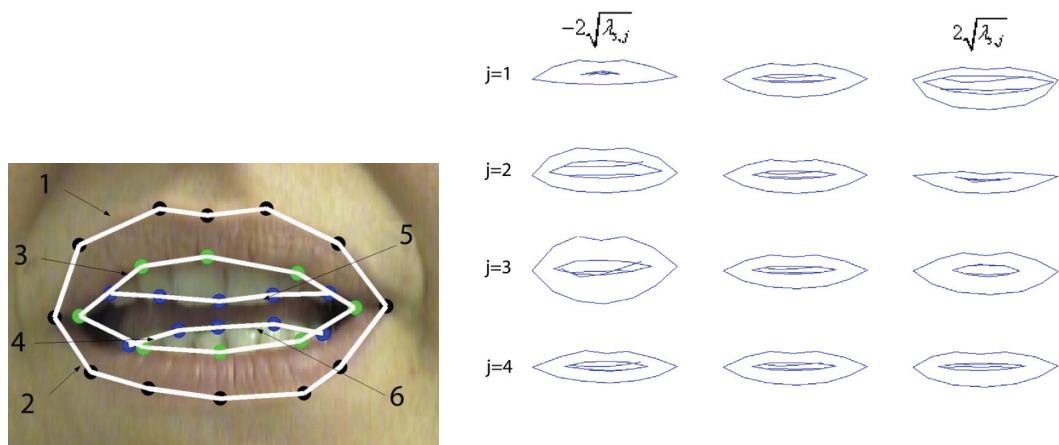


Figure 1.14 : Exemple de modèle de forme de bouche. A gauche, on donne l'ensemble des points clés utilisés pour échantillonner la forme de la bouche. A droite, on donne les 4 principaux modes de variations de la forme de la bouche.

Etant donné un vecteur x_i contenant les N points échantillonnant la forme de la bouche, une analyse en composantes principales (*ACP*) (en anglais, Principal Component Analysis (*PCA*)) est effectuée sur $X=[x_1, x_2, \dots, x_M]$, M étant le nombre d'images annotées, pour obtenir les modes de variations $P_s=[p_1, p_2, \dots, p_t]$. Soit la matrice $P_s=[p_1, p_2, \dots, p_t]$ des modes de

variations orthogonaux donnés par *PCA* et un vecteur de paramètres $b_s = [b_1, b_2, \dots, b_t]^T$, la forme de la bouche peut alors être décrite par la relation suivante : $x = \bar{x} + P_S b_S$, avec \bar{x} la forme moyenne de la bouche. Habituellement, 95% de la variance totale des données est conservée. Ceci amène alors à une réduction importante du nombre de dimensions du problème. La complexité du modèle est réduite car une forme particulière peut alors être décrite par un jeu de paramètres réduit.

Pour segmenter les lèvres sur une image inconnue, il faut déterminer le jeu de paramètres b qui minimisera une fonction de coût. Le rôle de la fonction de coût est de quantifier la différence entre la forme donnée par le modèle de forme actif et celle de l'image courante. Dans (Cootes, 2004), l'auteur propose une approche descendante pour optimiser un modèle de forme de visage. Le cas d'une application aux lèvres peut être abordé de manière identique. Pour chaque image de la base d'apprentissage, une pyramide gaussienne est construite. Puis, pour tout les points clés de l'image, le gradient est calculé le long de la normale au contour en ce point et ceci pour tous les niveaux de la pyramide. On obtient alors des modèles de distribution du gradient sur la normale au contour en chaque point clé. Les profils normaux considérés sont de même taille pour tous les niveaux de la pyramide. La procédure d'optimisation commence au niveau le plus haut de la pyramide, soit le niveau le moins détaillé. Les points clés sont initialisés sur le contour moyen. Ensuite, à chaque itération, la position des points est ajustée en minimisant la distance entre le profil du gradient sur la normale au contour et le modèle de profil correspondant à ce point clé. Les nouveaux paramètres de formes b_s' sont alors calculés à partir des nouvelles positions x' des points clés :

$$b' = P_s^{-1} (x' - \bar{x}) \quad (8)$$

La procédure est répétée jusqu'à la convergence du modèle de forme. Quand la convergence est atteinte au niveau L , l'optimisation est recommandée au niveau $L-1$ avec les paramètres de forme trouvés au niveau L jusqu'à atteindre $L=0$. $L=0$ correspond à la pleine résolution de l'image. Zhang utilise également un modèle de forme actif pour segmenter le visage, les yeux et la bouche (Zhang, 2003). Les modèles de gradient le long des normales aux contours aux positions des points clés sont remplacés par les réponses des filtres provenant d'une rosace de 40 filtres de Gabor. Le but est d'améliorer la robustesse de

l'optimisation. Dans (Jang, 2006) un algorithme utilisant un *ASM* est employé pour extraire les contours internes et externes de la bouche. Le modèle *ASM* est entraîné sur des images de bouche manuellement annotées. Des modèles normalisés des profils de la luminance pour chaque point clé sont construits. Les paramètres de forme pour les images de la base sont également calculés. Ensuite les auteurs modélisent la distribution des paramètres à l'aide d'un mélange de gaussiennes. La boucle est basée comme pour (Cootes, 2004) sur la minimisation la distance de Mahalanobis entre les profils de luminance courants aux points clés et les modèles. Les nouveaux paramètres de forme sont calculés à partir des nouvelles positions des points clés. En utilisant le modèle de distribution des paramètres, il y a alors réajustement si la probabilité de l'ensemble est trop faible. On peut également citer les travaux de (Nguyen, 2010) qui a développé un modèle statistique de forme hybride pour segmenter les contours externes et internes de la bouche. L'optimisation est basée sur 3 caractéristiques : des profils de luminance, des voisinages de luminance et les réponses à des ondelettes de Gabor au niveau des points clés des contours. Dans (Li, 2006) les auteurs proposent une autre variante pour optimiser le modèle de forme actif de la bouche. La texture autour des points clés est, cette fois-ci, modélisée par un classifieur Adaboost. Pour entraîner le classifieur, pour chaque image de la base d'apprentissage et pour chaque point clé du contour, on considère un voisinage 3x3 autour du point clé. Pour chaque point du voisinage 3x3, on prélève alors un échantillon 24x24 de la luminance centré sur ce point. Ces voisinages sont considérés comme des échantillons positifs. Ensuite, on prélève des échantillons de texture 24x24 dont le pixel central est positionné à l'intérieur d'un voisinage 12x12 autour du point clé, en excluant les points du voisinage 3x3 direct. Ils sont considérés comme des échantillons négatifs. Tous ces échantillons de texture sont décomposés sur une base d'ondelettes de Haar. Le classifieur AdaBoost est entraîné sur les données résultantes. Lors de l'optimisation du modèle *ASM*, un point clé sera déplacé vers la position qui maximise la confiance donnée par le classifieur Adaboost.

1.4.3.2 Modèles d'apparence actifs

Nous avons pu voir précédemment que pour l'optimisation des modèles de forme, l'information de texture est très souvent employée. Les modèles d'apparences actifs ou *AAM* (Active Appearance Model) qui ont également été proposés par Cootes (Cootes,

1998 ; 2000 ; 2001) ont été développés pour décrire conjointement les variations de forme et d'apparence. En partant des images où les contours de la caractéristique étudiée ont été annotés, Cootes propose d'échantillonner également la luminance de l'image (figure 1.15).

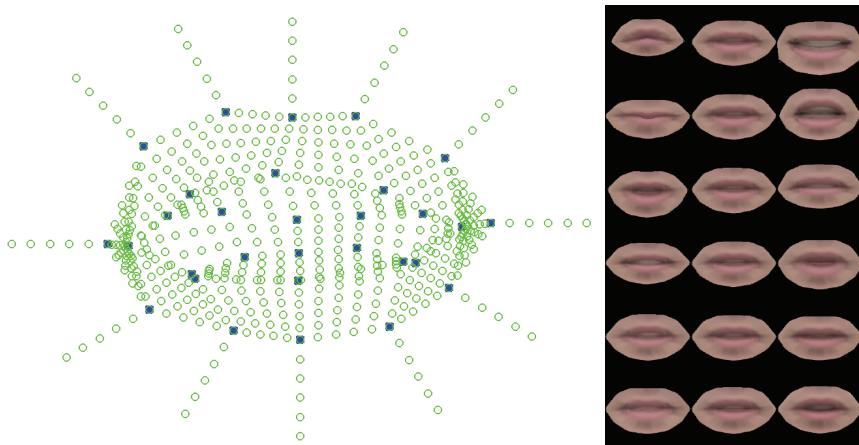


Figure 1.15 : Exemple de modèle d'apparence de bouche. A gauche, on donne la grille d'échantillonnage de la texture de la bouche. A droite, on donne les 6 premiers modes de variation du modèle d'apparence.

Pour cela, Cootes calcule la transformation qui permet de passer, pour chaque image, du contour courant au contour moyen. La transformation est ensuite appliquée à tous les pixels de la texture de l'image courante. Le vecteur contenant les échantillons de texture est ensuite normalisé pour éliminer les variations de lumière. Une analyse en composantes principales (*ACP*) est effectuée pour obtenir les modes de variations principaux de la texture. Comme pour la forme, l'apparence d'une image de bouche est donnée par la transformation linéaire $g = \bar{g} + P_g b_g$ avec \bar{g} l'apparence moyenne, P_g la matrice orthogonale des modes de variation de l'apparence et b_g le vecteur des paramètres de l'apparence (figure 1.15). La forme et l'apparence peuvent alors être décrites conjointement par un vecteur de paramètres d :

$$d = \begin{bmatrix} W_s P_s^T (x - \bar{x}) \\ P_g^T (g - \bar{g}) \end{bmatrix} \quad (9)$$

W_s est une matrice diagonale composée de coefficients de normalisation. En appliquant une nouvelle *ACP* sur l'ensemble des vecteurs d_i des images de la base, on obtient un modèle combinant l'apparence et la forme :

$$d=Q c \quad (10)$$

où Q est la matrice orthogonale des modes de variations donnée par *ACP* sur les paramètres combinés, c est un vecteur de paramètres contrôlant la forme et l'apparence. Des expressions détaillées sont données dans (Cootes, 2004).

Pour trouver les paramètres optimaux décrivant une image inconnue, une optimisation du modèle est effectuée. Le but est de minimiser la distance entre l'apparence de l'image et celle donnée par le modèle.

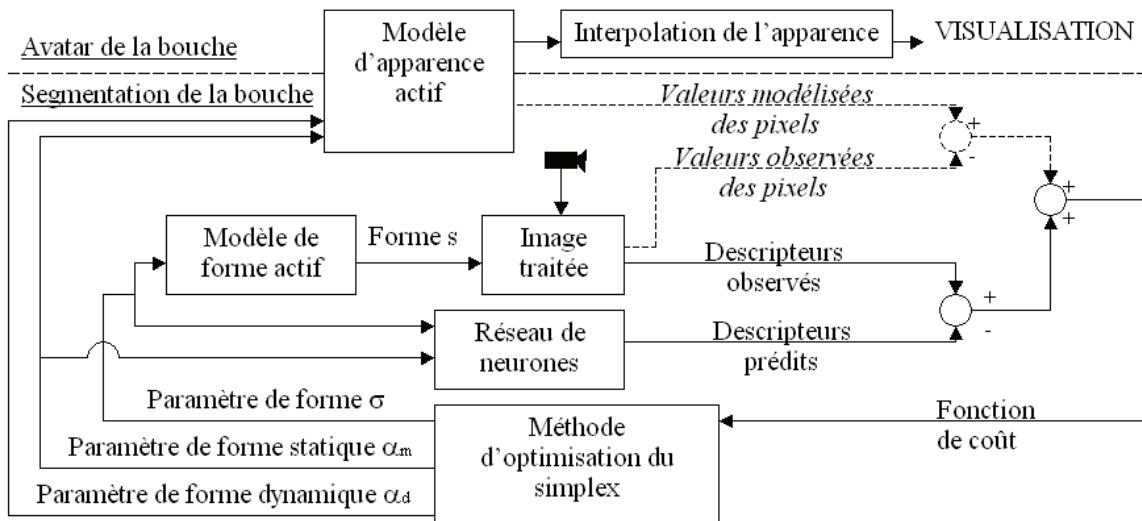


Figure 1.16 : Schéma de principe de l'algorithme proposé par (Gacon, 2005)

Gacon (Gacon, 2005) a proposé un modèle *AAM* multi-locuteur pour la segmentation des lèvres (figure 1.16). Dans ce travail, la base d'apprentissage est composée d'images de bouche de plusieurs sujets traités dans l'espace *YCbCr*. Pour gérer l'aspect multi-locuteur, Gacon propose 2 modèles d'apparence, un modèle décrivant l'apparence statique et un modèle décrivant l'apparence dynamique. Le modèle d'apparence statique est calculé à partir de l'apparence moyenne de chaque locuteur. Pour chaque locuteur, l'apparence moyenne est ensuite soustraite aux autres images. Une *ACP* est alors calculée sur ces apparences normalisées. Gacon calcule également les descripteurs G , G_x et G_y pour chaque point clé. G est un filtre gaussien, G_x et G_y sont les gradients verticaux et horizontaux de

l'image filtrée par G . Ces réponses sont utilisées pour entraîner un perceptron multicouche destiné à prédire les réponses de G , G_x et G_y à partir des paramètres de forme. L'optimisation de l'ensemble des paramètres est réalisée par une méthode du type descente du simplex. La fonction que l'on cherchera à minimiser inclut la distance entre l'apparence donnée par le modèle et celle de l'image ainsi que la distance entre les réponses de G , G_x et G_y de l'image aux positions des points clés et celles données par le réseau de neurones. La figure 1.17 donne des exemples des contours segmentés par la méthode (Gacon, 2005).



Figure 1.17 : Exemples de contours segmentés par l'algorithme présenté par Gacon (Gacon, 2005)

Dans (Matthews, 1998), une autre procédure est présentée pour optimiser un modèle *AAM* de bouche. Pendant la phase d'entraînement du modèle d'apparence, les paramètres d'apparence c sont calculés pour toutes les images de la base. Ensuite, pour minimiser l'erreur δE entre l'apparence générée par le modèle et celle de l'image courante, les auteurs calculent la transformation A , telle que $\delta c = A \delta E$ et où δc est l'erreur sur les paramètres. Pour estimer A , des erreurs aléatoires sont ajoutées aux paramètres d'apparence c . La transformation A est alors estimée par régression linéaire. L'optimisation du modèle *AAM* se déroule selon les étapes suivantes :

- On calcule l'erreur δE
- On calcule δc : $\delta c = A \delta E$
- On corrige les paramètres : $c' = c - \delta c$
- Une nouvelle apparence est calculée à partie de c'
- Les étapes précédentes sont répétées jusqu'à ce que l'erreur soit inférieure à un seuil prédéfini.

L'hypothèse, faite dans (Matthews, 1998), qu'il existe une relation linéaire entre δE et δc n'étant pas réaliste, Matthews propose une méthode d'optimisation modifiée pour tenir compte de la non-linéarité. L'optimisation est réalisée par une méthode du type descente du gradient où les modèles de forme et d'apparence ne sont plus combinés. Une étape d'optimisation de la forme est effectuée en premier. Ensuite, on optimise l'apparence par la descente du gradient.

L'intérêt principal des méthodes du type *ASM* et *AAM* est que le résultat de la segmentation sera toujours réaliste, car la forme et l'apparence sont entraînées sur des exemples réels. Bien entendu, comme pour les autres types de méthodes supervisées, cela implique la segmentation manuelle d'un nombre conséquent d'exemples de manière à avoir une base la plus exhaustive possible. De plus, comme l'a montré Yang dans son étude (Yang, 1996), un changement de système d'acquisition, même en environnement contrôlé, introduit des changements importants sur les propriétés des images (rapport signal-sur-bruit, rendu des couleurs, ...). Une étape de normalisation entre les dynamiques des images de la base d'apprentissage est alors nécessaire. L'autre problème, dans le cas des modèles *AAM*, est que l'hypothèse de linéarité entre les paramètres de forme et d'apparence n'est pas pertinente. La convergence du modèle vers un minimum global ne sera pas garantie. La performance de l'algorithme dépendra alors de la qualité de l'initialisation. Plus elle sera proche du résultat désiré, plus l'hypothèse de linéarité sera pertinente.

1.5 Segmentation des lèvres : Approches Contour

Les approches « contour » pour la segmentation des lèvres sont essentiellement basées sur des modèles de contours déformables. Ce type de méthode consiste à choisir et à optimiser un modèle mathématique (exemple : une spline, un polynôme, ...) pour qu'il s'adapte aux contours de l'objet d'intérêt. La déformation du modèle initial est guidée par la minimisation d'une fonction d'énergie. Cette fonction est composée d'un terme d'énergie interne, décrivant les propriétés géométriques du contour, et d'un terme d'énergie externe, calculé à partir des propriétés de l'image. Les modèles déformables sont divisés en deux grandes catégories : les contours actifs et les modèles paramétriques. Les contours actifs modélisent les contours d'un objet d'intérêt à partir d'une chaîne initiale de points, dont on

modifie la position successivement. Ces modèles sont très flexibles, car ils n'incluent pas de contraintes a priori sur la forme des objets que l'on cherche à modéliser. Les modèles paramétriques définissent une description de l'objet d'intérêt intégrant une connaissance a priori sur la forme globale et contraignant la forme finale du contour.

1.5.1 Contours actifs

Les contours actifs ou « snakes », introduit par Kass et al. (Kass, 1987), sont, dans notre cas, composés d'un ensemble de points mobiles localisés sur une courbe située dans le plan. Suivant l'application, la courbe peut être fermée ou ouverte, avec des extrémités fixes ou mobiles. Les snakes évoluent itérativement, depuis une position initiale, jusqu'à la position finale qui minimise une fonction d'énergie. Dans le contexte de la segmentation des lèvres, les snakes sont largement utilisés, car ils permettent d'extraire des contours complexes.

1.5.1.1 Définition et propriétés des contours actifs

Un snake est représenté dans le plan par une courbe $\nu(s) = (x(s), y(s))$, s étant l'abscisse curviligne, est par sa fonction d'énergie qui est définie de la manière suivante :

$$\varphi(\nu) = \int (E_{\text{int}}(\nu(s)) + E_{\text{ext}}(\nu(s))) ds \quad (11)$$

L'énergie interne E_{int} décrit les propriétés mécaniques de la courbe. Les forces internes imposent des contraintes locales sur les points de la courbe. Elles ont une influence limitée sur la forme globale du contour final, mais elles permettent une régularisation de la courbe lors de la déformation. L'énergie externe E_{ext} est liée aux propriétés de l'objet d'intérêt. Elle sera minimisée par déformation de la courbe vers les zones saillantes de l'objet, telles que les contours. Etant donné l'absence de contrainte sur la forme de l'objet à segmenter, les contraintes internes du snake devront être assez importantes pour empêcher la courbe de diverger. Généralement, le terme d'énergie interne est de la forme suivante :

$$E_{\text{int}}(\nu) = \frac{1}{2} \int (\alpha(s)|\nu'(s)|^2 + \beta(s)|\nu''(s)|^2) ds \quad (12)$$

où $v'(s)$ et $v''(s)$ correspondent aux dérivées première et seconde de $v(s)$, $\alpha(s)$ est le coefficient d'élasticité de la courbe. Ce coefficient aura une influence sur la tension de la courbe : une valeur importante de α conduira à des courbes avec peu de discontinuités locales tandis qu'une valeur faible autorisera des discontinuités locales importantes. $\beta(s)$ représente le coefficient de rigidité qui influence la courbure. Des valeurs faibles de β autoriseront des courbures importantes. Inversement, des β importants limiteront la courbure du snake. L'énergie externe du snake est calculée à partir des caractéristiques de l'image. Par exemple, pour extraire les contours, le gradient est couramment employé. La luminance peut également être utilisée pour certaines applications. L'optimisation du snake est réalisée par déplacement des points de la courbe, jusqu'à minimisation de l'énergie du snake.

Les contours actifs sont largement utilisés pour les problèmes d'extraction de contour de part leur flexibilité et leur simplicité. Les snakes peuvent être utilisés indifféremment pour extraire des contours ouverts ou fermés, sans contraintes de forme. L'optimisation revient à itérer un système linéaire stable au regard des contraintes internes sur la courbe. La qualité de l'optimisation va grandement dépendre des paramètres du snake : de mauvais réglages des paramètres pourront, par exemple, empêcher le snake de converger sur les concavités des contours d'intérêt, ou le faire converger vers des contours qui n'existent pas. Le réglage de ces paramètres est difficile, si bien que, la plupart du temps, ils sont fixés de manière empirique. Plusieurs méthodes ont été proposées pour lever ces difficultés. Cohen (Cohen, 1991) a introduit le terme d'énergie ballon pour forcer le snake à se contracter, tandis que le flux du gradient est utilisé dans (Xu, 1998). Ballerini a introduit le concept de snakes génétiques pour améliorer l'optimisation (Ballerini, 1999). Pour le cas particulier de la segmentation des lèvres, les snakes ont été également employés. Dans la suite, nous donnerons un aperçu des principaux travaux existants en nous concentrant, en premier lieu, sur les différents types de snakes qui ont été utilisés. Nous reviendrons aussi sur les principales méthodes d'initialisation, avant de nous concentrer sur les principales fonctions d'énergie qui ont été introduites pour déformer les snakes.

1.5.1.2 Contours actifs pour la segmentation labiale

Différents types de contours actifs ont été proposés pour segmenter la bouche, avec pour objectif de réduire les problèmes de dépendance à l'initialisation et d'améliorer la convergence.

Wu et al. (Wu, 2002) ont introduit un contour actif utilisant un champ de gradients composites (*GVF snakes*, Gradient Vector Flow snakes). L'utilisation du *GVF* permet d'initialiser le snake loin de l'objet à segmenter. La principale limitation de cette méthode est liée au coût important en temps de calcul du calcul du *GVF*. Son utilisation dans des applications temps-réel paraît difficile.

Dans (Séguier, 2003), des « snakes génétiques » sont introduits. Dans cette implémentation, l'optimisation est réalisée à l'aide d'une méthode de type génétique afin de réduire le risque de convergence vers un minimum local de la fonction d'énergie. L'inconvénient de ce type de méthode d'optimisation réside également dans le temps de calcul. Une méthode de type génétique demande un grand nombre de réalisations pour trouver la famille de paramètres optimale.

Shinchi propose un contour actif échantillonné (*SCAM*, Sampled Active Contour Model) pour modéliser le contour externe des lèvres (Shinchi, 1998). Dans cette implémentation, les forces internes et externes sont appliquées sur les points de contrôle plutôt que d'être intégrées sur toute la courbe. Cela permet d'accélérer la convergence du snake vers le contour externe des lèvres. L'inconvénient de cette méthode vient de l'application des contraintes sur les points de contrôle uniquement, ce qui paraît peu robuste.

Dans (Kaucic, 1998) et (Wakasugi, 2004), les auteurs utilisent des B-snakes pour segmenter le contour des lèvres. Originalement développés par (Blake, 1995), ces snakes utilisent des B-splines pour calculer la courbe à partir des points de contrôle. Les propriétés intrinsèques des splines permettent de ne considérer que le terme d'énergie externe lors de la déformation de la courbe. Le coût en temps de calcul est également réduit par rapport aux snakes classiques.

1.5.1.3 Contours actifs : Initialisation

L'initialisation d'un snake est une étape clé si l'on veut éviter la convergence du snake vers un minimum local. La courbe initiale devra donc être la plus proche possible du

contour final. De plus, les snakes ayant tendance à se contracter, ils doivent être initialisés à l'extérieur de la bouche. Le choix des paramètres du snake résulte d'un compromis, ils doivent contraindre suffisamment le snake pour qu'il ne soit pas sensible aux contours parasites, mais ne pas empêcher la convergence vers les zones désirées. Pour la segmentation des lèvres, plusieurs approches ont été proposées pour initialiser les contours actifs.

- *Extraction de la zone d'intérêt de la bouche*



Figure 1.18 : a) Accumulation verticale des pixels sombres, b) Projection horizontale de l'intensité du gradient de la luminance (Delmas 1999).

Le but, à cette étape, est d'obtenir les coordonnées d'un rectangle englobant la bouche. Ce rectangle servira de courbe initiale. Des méthodes basées sur la couleur, permettant de réaliser cette opération, ont été présentées à la section 1.3. D'autres méthodes utilisent les gradients d'intensité pour localiser la zone de la bouche.

Dans (Radeva, 1995), la position de la bouche est déterminée en analysant les sommes cumulatives de l'image horizontalement et verticalement. La position de la bouche est ensuite recherchée en analysant ces « projections » sur l'horizontale et la verticale et en ajoutant une contrainte de symétrie. Dans (Delmas, 1999; Delmas, 2002; Kuo, 2005) la position verticale de la bouche est déterminée par accumulation des pixels les plus sombres de chaque colonne de l'image (figure 1.18-a). Les commissures des lèvres sont ensuite positionnées en chaînant les pixels les plus sombres en partant du centre de la bouche, et en détectant les sauts. La projection de l'intensité du gradient selon l'horizontale donne les frontières, haute et basse, de la bouche (figure 1.18-b).

- *Contour initial*

Dans (Seo, 2003; Kuo, 2005), le contour initial est simplement placé sur la boîte englobant la bouche. Dans (Delmas, 1999), l'auteur tente d'identifier les frontières internes et externes des lèvres en analysant la carte des contours de la zone de la bouche. Des points de contrôle sont ensuite positionnés sur les zones identifiées. D'autres auteurs utilisent des modèles introduisant des caractéristiques morphologiques de la bouche. Dans (Chiou, 1997), l'auteur place un cercle centré sur le barycentre de la zone de la bouche. Des points de contrôle sont ensuite régulièrement placés sur le cercle. Leurs positions sont ajustées en faisant varier la norme du vecteur les reliant au centre du cercle. Dans (Delmas, 2002), les positions estimées des extrema de la bouche sont utilisées pour calculer un contour composé de quadriques. Ce contour est ensuite échantillonné pour obtenir les points de contrôle du snake. Beaumesnil calcule des courbes cubiques à partir des extrema estimés de la bouche pour initialiser le contour externe (Beaumesnil, 2006). Pour initialiser le contour interne, Beaumesnil applique une transformation anisotrope sur le snake externe, après convergence, par rapport au barycentre des lèvres et à l'estimation de leur épaisseur. Une méthode similaire est proposée dans (Seyedarabi, 2006), l'auteur utilise des ellipses comme contours initiaux.

Dans le cas d'algorithme de suivi, la première image de la séquence est initialisée avec une des méthodes précédentes. Par la suite, la position du contour initial sur l'image au temps t est déterminée en utilisant directement le contour extrait à $t-1$ (Seyedarabi, 2006), ou en le dilatant (Kuo 2005). Dans (Delmas 2002 ; Beaumesnil 2006), un suivi des points de contrôle est effectué entre chaque image. Les nouvelles positions des points de contrôle servent alors d'initialisation. D'autres auteurs utilisent des techniques de reconnaissance de patrons (template matching) pour initialiser le snake à partir de l'image précédente (Barnard, 2002; Wu, 2002; Seo, 2003).

1.5.1.4 Contours Actifs : Fonctions d'énergie

Une fois le contour initial déterminé, l'algorithme d'optimisation du contour doit déformer le snake de manière à le faire converger vers le contour des lèvres par minimisation de la fonction d'énergie du snake. De l'expression de la fonction d'énergie dépendra la

convergence du snake. Plusieurs formulations des termes de la fonction d'énergie (E_{int} et E_{ext}) du snake ont été proposées pour la segmentation des lèvres.

- *Energie interne et contrainte locale du snake*

Le terme E_{int} modélise l'influence des forces internes du snake. Les forces internes imposent des contraintes locales sur les points de la courbe. Elles auront tendance à « lisser » le contour. En théorie, les coefficients α et β sont variables (cf. (12)), mais généralement ils sont fixés empiriquement et considérés constants.

Le fait que le modèle de contour actif ne prenne pas en compte des contraintes de formes peut également créer des difficultés. Par exemple, le cas des commissures de la bouche est problématique. Ces zones sont, très souvent, mal définies, floues, avec des gradients faibles. De plus, la forte courbure du contour au niveau des commissures rendra difficile la convergence d'un snake. Des contraintes locales doivent être ajoutées pour modéliser au mieux le contour au niveau des commissures.

Dans (Seyedarabi, 2006), les points de contrôle initiaux du snake sont régulièrement espacés, sauf autour des commissures où la densité de points de contrôle est plus élevée. Dans (Delmas, 1999; Pardas, 2001; Seguier, 2003, Kuo 2005), le coefficient β n'est pas considéré constant. Sa valeur est grande au milieu de la bouche et tend vers zéro lorsqu'on se rapproche des commissures. L'hypothèse est que la courbure du contour est faible au milieu de la bouche et maximale au niveau des commissures.

D'autres contraintes peuvent être ajoutées par l'intermédiaire de E_{int} . Par exemple, dans (Radeva, 1995), l'auteur ajoute une contrainte de symétrie du contour externe par rapport à la verticale passant par le centre de la bouche.

Pour réduire l'influence des paramètres α et β et pour régulariser le snake pendant la convergence, des auteurs ont proposé d'utiliser des modèles de bouche. Wu et al. ajoutent un terme à l'énergie interne dans (Wu, 2002). Tout d'abord, les auteurs optimisent un modèle de contour externe composé de 2 paraboles sur la carte des contours obtenue par un détecteur de Canny. Un filtrage passe-bas est appliqué au modèle optimisé sur la carte des contours. Le champ produit par le filtrage passe-bas est ajouté à E_{int} . Ce terme aura pour effet d'ajouter une contrainte de forme lors de la convergence du snake.

- Energie externe

Les forces externes sont calculées à partir de l'image. Elles déformeront le snake vers les zones saillantes de l'image lors de la convergence. Dans le cas de la bouche, nous sommes intéressés par les contours. Dans (Seyedarabi, 2006), la force externe est définie comme la somme de la luminance de l'image et du gradient de luminance. Radeva remarque que la convergence d'un snake est plus difficile pour le contour inférieur externe de la bouche, à cause des variations de lumière qui peuvent rendre cette zone inhomogène (Radeva, 1995). Il peut exister des réflexions spéculaires et des ombres suivant la direction de la source lumineuse. Pour résoudre ce problème, Radeva définit trois snakes avec des expressions de la force externe adaptées aux trois conditions d'illumination suivantes : lèvre inférieure sombre et peau claire, lèvre claire et peau sombre, lèvre incluant des zones sombres et claires. Le snake, dont l'énergie après convergence est minimale, est choisi pour modéliser le contour inférieur. Dans (Wu, 2002), la force externe correspond à une expression particulière du gradient. Le champ de gradient construit peut être vu comme le résultat de la diffusion du champ de gradient de la luminance ou du gradient d'une image binaire. L'intérêt de cette diffusion est d'augmenter la force d'attraction des contours d'intérêt. Le snake sera alors moins dépendant de l'initialisation car il pourra être initialisé loin de l'objet. L'inconvénient majeur de cette méthode est la résolution des équations générales de la diffusion qui est très exigeante en temps de calcul. Cela rendra cette implémentation difficilement utilisable pour des applications temps-réel. Pardas ajoute un terme à la force externe en prenant en compte l'information donnée par l'image précédente dans le cas de séquences vidéo (Pardas, 2001). Ce terme diminue la force externe quand les zones autour des points de contrôle se ressemblent d'une image à l'autre.

L'information de couleur peut être également utilisée pour calculer la force externe. Beaumesnil propose de combiner la teinte et la luminance pour calculer le gradient qui est utilisé comme force externe (Beaumesnil, 2006). Seo, dans (Seo, 2003), modélise la couleur à l'intérieur et à l'extérieur du snake par des distributions gaussiennes sur les composantes R et G de l'espace RGB . Le snake est alors déformé pour maximiser la distance entre les distributions de couleur interne (pixels des lèvres) et externe (pixels de peau). Kaucic applique une analyse discriminante linéaire sur les valeurs de teinte des

pixels de peau et des lèvres pour augmenter le contraste (Kaucic, 1998). Le gradient calculé est utilisé comme force externe. Dans (Chiou, 1997), le masque des lèvres, obtenu par seuillage de $Q=R/G$, est utilisé pour calculer le terme principal de la force externe. Lors de la déformation du snake, la valeur de l'énergie externe dépendra du nombre de pixels du masque se trouvant dans les voisinages entourant les points du contour.

Seguier définit une fonction d'énergie globale pour optimiser simultanément deux snakes, un pour le contour extérieur des lèvres, et un pour le contour interne de la bouche (Seguier, 2003). Le premier terme de la fonction contrôle la rigidité des deux snakes. Pour cela, l'auteur calcule pour chaque point de contrôle des snakes la somme des différences absolues entre le niveau dans V (de l'espace couleur YUV) du point de contrôle et les niveaux dans V de ses voisins directs. Pour le deuxième terme, un masque binaire de l'intérieur de la bouche est obtenu par seuillage sur V . L'auteur détermine ensuite le nombre N de pixels du masque entourés par le snake modélisant le contour interne. Seguier affirme que les pixels des lèvres ont des niveaux élevés dans V . Le troisième terme de la fonction d'énergie est construit de manière à être petit quand la somme des niveaux des pixels dans V entre les deux snakes est grande et que le nombre de pixels est petit.

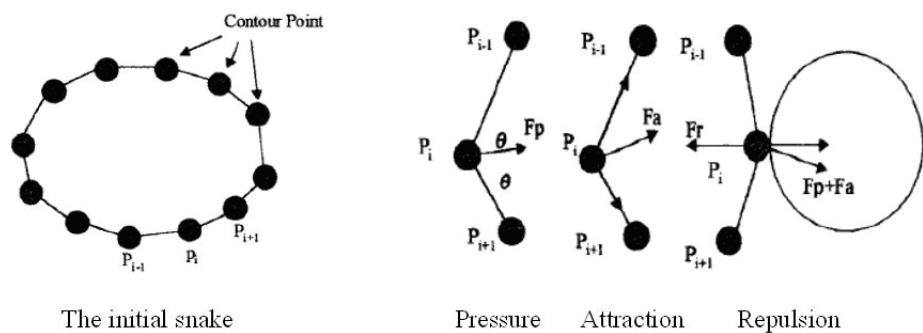


Figure 1.19 : Les 3 forces externes F_p , F_a et F_r proposées par (Schinchi, 1998).

Dans (Shinchi, 1998), les contraintes externes résultent de 3 forces. Une force de pression F_p et une force d'attraction F_a contractent le snake vers le contour. Lorsque le snake rencontre une zone de contour, une force de répulsion F_r s'oppose aux 2 forces précédentes (figure 1.19). Un facteur de vibration du contour est appliqué pour permettre au snake

d'éviter les contours parasites. Cette force est parallèle à la somme des forces d'attraction et son sens est inversé à chaque itération pendant la déformation. Enfin dans (Barnard, 2002), la minimisation de l'énergie est guidée par appariement de blocs avec des modèles des voisinages des points de contrôle du snake.

Dans le cas des contours fermés, une force externe additionnelle, dite ballon, a été proposée par Cohen (Cohen, 1991). La force ballon a pour effet de contracter ou de dilater le snake comme un ballon. Le but est d'éviter les minimums locaux de la fonction d'énergie. Initialement, Cohen a proposé une force dont l'effet est de dilater un snake qui serait initialisé à l'intérieur du contour à extraire, comme dans (Chiou, 1997). Cette force peut aussi être utilisée pour compenser la tendance du snake à se contracter (Delmas, 1999). Dans la pratique, ce type de force est difficile à utiliser. Si la force est trop grande, le contour se dilatera jusqu'aux limites de l'image ou se contractera en un point. Si elle est trop faible, le contour n'évitera pas les minimums locaux. Kuo en 2005 a défini une force homogène à une pression. Cette force est calculée à partir de la couleur à l'intérieur du contour (Kuo, 2005). La force ballon gonfle ou dégonfle le snake suivant l'homogénéité des teintes des régions segmentées par le snake. Beaumesnil initialise le snake à l'extérieur de la zone de la bouche (Beaumesnil, 2006). Une force ballon est appliquée pour contracter le snake vers le barycentre de la zone de bouche.

1.5.1.5 Discussion

Le problème majeur rencontré lors de l'utilisation des contours actifs pour la segmentation des lèvres est l'initialisation. La convergence vers un minimum global de la fonction d'énergie n'est pas garantie. Si l'initialisation est grossière, le snake peut converger vers des minimums locaux de la fonction d'énergie. La robustesse aux changements des conditions de l'environnement (éclairage, sujet) avec ce type d'approche est également incertaine à cause du nombre de paramètres à définir. Il faut également noter que la plus grande partie des contributions se concentre sur l'extraction du contour externe de la bouche. En effet, le contour interne est un problème plus complexe que le contour externe. L'apparition et la disparition des dents, de la langue et de la cavité buccale, conduit à un nombre important de contours parasites qui peuvent perturber la segmentation. La modélisation des commissures des lèvres par des snakes est également un problème

difficile. Ces zones sont, la plupart du temps, sombres et avec un contour présentant une forte courbure au niveau des commissures. Les forces externes sont souvent trop faibles pour compenser les contraintes de rigidité des snakes et déformer correctement le snake au niveau des commissures.

Néanmoins, les contours actifs représentent des solutions intéressantes en fin d'une chaîne de traitement visant à la segmentation des lèvres grâce à leur grande flexibilité. La simplicité et la vitesse de convergence des snakes sont également des avantages, notamment pour les applications de suivi de bouche.

1.5.2 Modèles paramétriques

Les modèles paramétriques ont de grandes similarités avec les modèles de contours actifs pour ce qui est de l'optimisation. La déformation, comme pour les snakes, obéit à la minimisation d'une fonction d'énergie. La grande différence réside dans le fait qu'une hypothèse sur la forme du contour recherché est explicitement faite dans leur définition.

1.5.2.1 Modèles paramétriques : Définition et propriétés

Les modèles paramétriques, introduits par Yuille et al. (Yuille, 1992), décrivent un objet à l'aide d'un patron ou template. La plupart du temps, ce patron est composé de différents types de courbes (cercles, ellipses, polynômes, splines, ...). Les patrons sont, comme les snakes, déformés dynamiquement par la minimisation d'une fonction d'énergie. La fonction d'énergie globale est composée d'un terme d'énergie interne et d'un terme d'énergie externe. Là où le terme d'énergie interne décrit les contraintes locales sur le snake, le terme d'énergie interne d'un modèle paramétrique décrit les variations possibles du patron. La liberté des modèles paramétriques sera limitée, au contraire des snakes, car la déformation du modèle sera guidée par un vecteur de paramètres, et non directement par le déplacement des points de contrôle. Ceci permet de gérer implicitement les occultations.

Le principal avantage apporté par les modèles paramétriques est l'ajout d'une contrainte de forme sur le contour que l'on souhaite modéliser. On évitera ainsi que le modèle converge vers des contours incompatibles avec la caractéristique étudiée. Toutefois, il faut avoir à l'esprit que ce gain potentiel en robustesse se fera au détriment de la précision dans le cas de déformations importantes de la bouche.

La définition d'un modèle paramétrique sera le résultat d'un compromis entre la déformabilité et les contraintes que l'on souhaite imposer au modèle. Plus on contraindra la forme du contour, moins le modèle sera en mesure de s'adapter à de nouvelles formes.

De manière générale, on peut mettre en avant trois grandes étapes dans la conception d'un algorithme de segmentation de contour basé sur un modèle paramétrique : la définition du patron des lèvres qui inclut une connaissance sur la forme recherchée, l'étape d'initialisation du contour sur une image inconnue et enfin la boucle d'optimisation du contour. Dans la suite, nous présenterons pour chacune de ces étapes, les différentes contributions en relation avec la segmentation des contours des lèvres.

1.5.2.2 Modèles paramétriques : Définition des modèles de contours

La bouche est un élément du visage qui peut présenter des variations de forme très importantes d'un sujet à un autre et au cours du temps (figure 1.20).



Figure 1.20 : Exemples d'images de bouche (Martinez, 1998)

Les modèles paramétriques sont généralement composés de courbes définies par morceaux. Chaque portion de la courbe peut alors être indifféremment définie par des polynômes, des splines, ou toutes autres fonctions suivant la précision et la complexité recherchée. Par exemple, pour le contour supérieur des lèvres, l'arc de cupidon (le V au milieu du contour supérieur) bénéficie souvent d'une modélisation particulière par des lignes brisées. La distinction est aussi faite entre le contour interne et le contour externe des lèvres. Des patrons peuvent être définis suivant l'état de la bouche (ouverte ou fermée) (Zhang, 1997; Yin, 2002; Stillittano, 2008) ou par rapport à la forme (légèrement ouverte, grande ouverte, ...) (Tian, 2000). En premier lieu, nous étudierons les modèles de contours internes avant de nous concentrer sur les modèles de contours externes de la bouche.

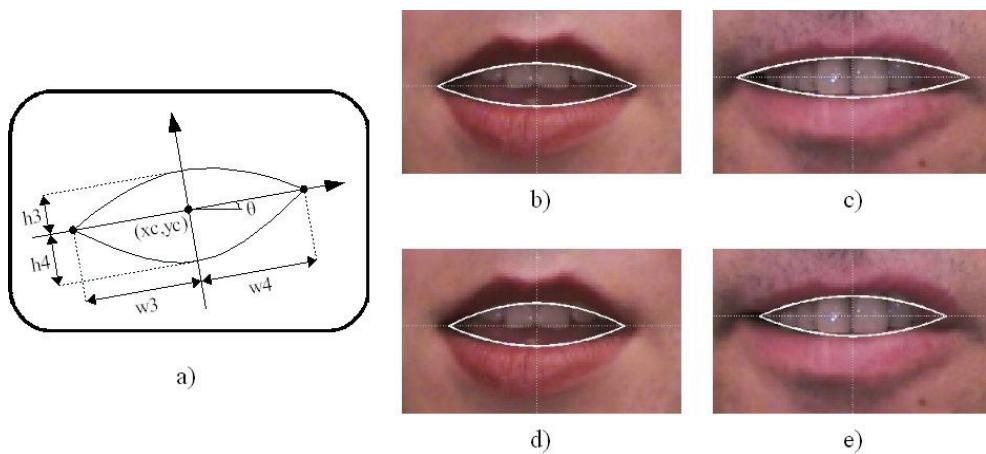


Figure 1.21 : Modèle de contour interne à 2 paraboles, a) Modèle de contour interne, sur les exemples b) et c), les paraboles sont jointes aux commissures externes de la bouche, sur les exemples d) et e) les paraboles sont jointes aux commissures internes de la bouche.

- Contour interne de la bouche

Deux considérations sont à prendre en compte, lors de la construction d'un modèle de contour interne de la bouche. Est-ce que ce modèle sera valable pour une bouche ouverte et une bouche fermée ? Est-ce que les commissures seront confondues avec celles du contour externe ou non ?

Le modèle classique de contour interne est composé de deux paraboles. On peut définir simplement ces paraboles par l'expression suivante :

$$y = h \left(1 - \frac{x^2}{w^2}\right) \quad (13)$$

où h représente la hauteur de la parabole et w la largeur. Dans (Yuille, 1992), le contour intérieur est modélisé par deux paraboles qui se rejoignent aux commissures externes des lèvres. Si la bouche est fermée, les deux paraboles sont confondues. Ce modèle impose une symétrie verticale qui peut ne pas être vérifiée (figure 1.21-c). Les cinq paramètres à optimiser sont (figure 1.21-a) : les coordonnées du centre (xc, yc) , l'inclinaison θ , la largeur de la bouche $w=w3+w4$, la hauteur $h3$ du contour supérieur et la hauteur $h4$ du contour inférieur. Le même modèle est employé par (Hennecke, 1994; Coianiz, 1996; Zhiming, 2002), à la différence que les paraboles se rejoignent aux commissures internes de la bouche (figure 1.21-d et 1.21-e). Zhang (Zhang, 1997) et Yin (Yin, 2002) distinguent les

cas « bouche ouverte » et « bouche fermée ». Quand la bouche est fermée, le contour interne est composé d'une unique parabole (figure 1.22). Le modèle à 2 paraboles est utilisé pour les bouches ouvertes. Wu (2002) emploie un modèle à 2 paraboles contrôlé par 4 paramètres : la distance haute, la distance basse, la distance à droite et la distance à gauche entre le contour externe et le contour interne de la bouche.

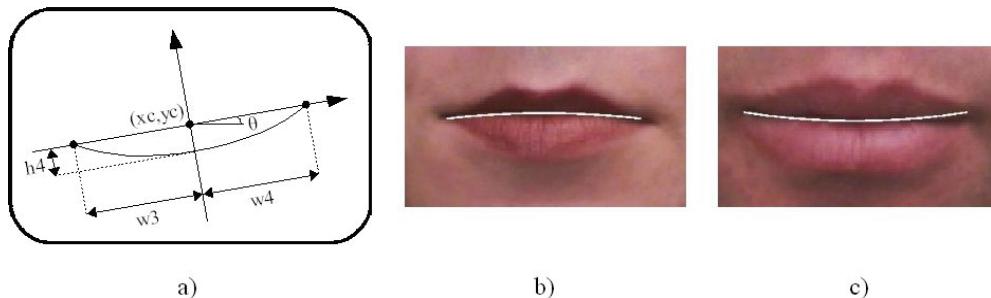


Figure 1.22 : Modèle de contour interne pour une bouche fermée, a) Modèle de contour interne à une parabole, b) et c) exemples de convergence du modèle.

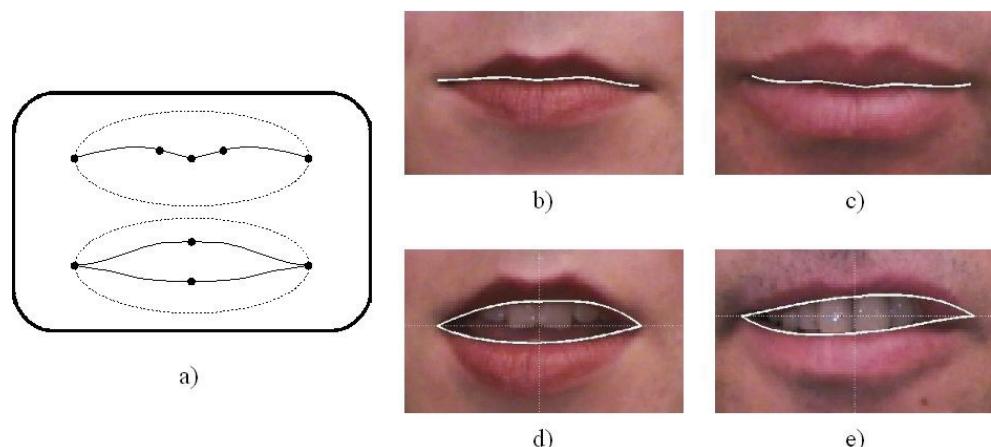


Figure 1.23 : Modèles de contours internes (Stillittano, 2008), a) Modèles paramétriques pour le contour interne des lèvres (Stillittano, 2008), b) et c) exemples de convergence pour des bouches fermées, d) et e) exemples de convergence du contour pour des bouches ouvertes.

Chen (Chen, 2006) s'affranchit de la contrainte de symétrie verticale pour le contour interne haut uniquement, w_3 n'est plus obligatoirement égal à w_4 . Pantic s'affranchit de la contrainte de symétrie verticale pour les deux contours (Pantic, 2001). Chaque moitié de la bouche est alors traitée séparément, ce qui amène à des contours internes asymétriques tout en gardant une complexité relativement faible (six paramètres au lieu de cinq avec (Yuille, 1992)). Dans (Stillittano, 2008), on distingue également les bouches ouvertes et fermées.

Pour les bouches ouvertes, le modèle de contour interne est construit avec 4 courbes cubiques (figure 1.23-a). Pour estimer chacune des courbes, cinq points sont nécessaires : une commissure, le centre du contour interne (haut ou bas), et trois autres points sur le contour. Le contour pour une bouche fermée est composé de deux courbes cubiques et de deux droites au centre (figure 1.23-a). Pour les deux cas, ouvert et fermé, les courbes se rejoignent aux commissures externes de la bouche (figure 1.23-b et 1.23-c).

Vogt ne calcule le contour que lorsque la bouche est fermée (Vogt, 1996). Le contour est calculé avec une courbe de Bézier à l'aide de six points. Malciu utilise des splines pour calculer les contours internes supérieur et inférieur en considérant cinq points pour chaque courbe (Malciu, 2000).

- Contour externe de la bouche

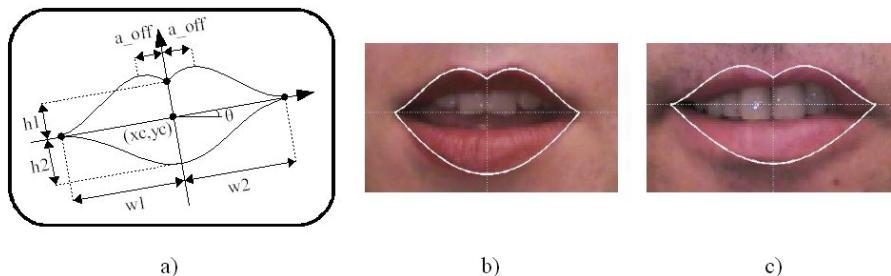


Figure 1.24 : Modèle de contour externe, a) Modèle paramétrique composé de 3 quadriques (Yuille, 1992), b) et c) exemples de convergence du modèle.

En ce qui concerne le contour extérieur de la bouche, les contours supérieurs et inférieurs doivent être définis différemment à cause de la présence de l'arc de Cupidon sur le contour supérieur. Le premier modèle paramétrique proposé pour la bouche, par Yuille (Yuille, 1992), est composé de 3 quadriques (figure 1.24) de la forme suivante:

$$y = h\left(1 - \frac{x^2}{w^2}\right) + 4q\left(\frac{x^4}{w^4} - \frac{x^2}{w^2}\right) \quad (14)$$

où h est la hauteur de la courbe, w est la largeur de la courbe, q détermine la déviation de la courbe par rapport à une parabole. Une contrainte de symétrie verticale est imposée au contour global. Dans le modèle proposé par Yuille, il y a huit paramètres à optimiser : les coordonnées du centre de la bouche (xc, yc), l'inclinaison θ , la largeur de la bouche

$w=w_1+w_2$ (avec $w_1=w_2=w_3=w_4$), les hauteurs h_1 et h_2 , le décalage a_{off} et le paramètre q . Le modèle complet, contour interne et contour externe, a au total 11 paramètres. La symétrie étant une hypothèse très restrictive, d'autres auteurs utilisent un modèle identique mais suppriment la contrainte de symétrie verticale : Dans (Hennecke, 1994), $w_1=w_2$ et $w_3=w_4$ mais $w_2 \neq w_3$. Dans (Yokogawa, 2007), l'auteur utilise 4 quadriques, 2 pour le contour externe haut et 2 pour le contour externe bas.

Plusieurs contributions proposent l'utilisation de paraboles pour définir le modèle paramétrique de la bouche (Rao, 1995; Zhang, 1997; Tian, 2000 ; Yin, 2002) (figure 1.25-a). Six paramètres sont alors à estimer : les coordonnées du centre de la bouche (x_c, y_c), l'inclinaison θ , la largeur de la bouche w_1+w_2 (avec $w_1=w_2$) , la hauteur h_1 et la hauteur h_2 . Les figures 1.25-b et 1.25-c montrent que ce modèle aboutit à une modélisation grossière du contour de la bouche.

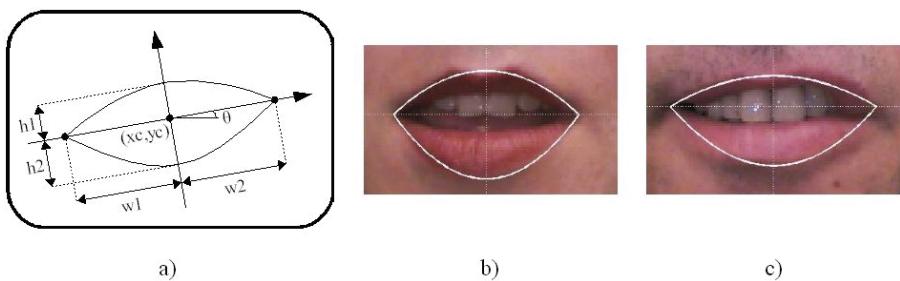


Figure 1.25 : Modèle de contour externe à 2 paraboles, a) Modèle paramétrique de bouche à 2 paraboles, b) et c) Exemples de convergence.

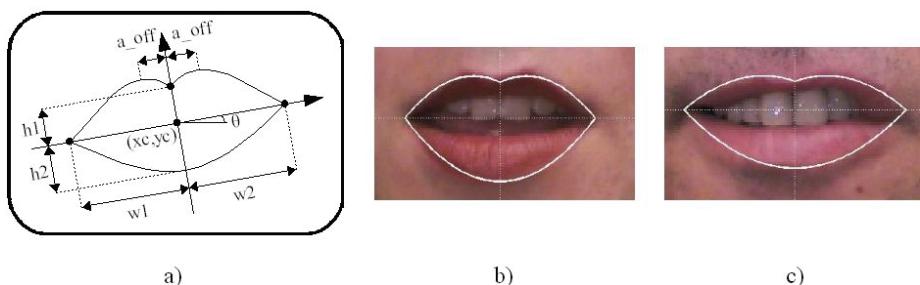


Figure 1.26 : Modèle de contour externe à 3 paraboles, a) Modèle paramétrique à 3 paraboles, b) et c) Exemples de convergence.

Liew (Liew, 2000) et Werda (Werda, 2007) appliquent une série de transformations géométriques pour affiner les contours donnés par le modèle à deux paraboles. Pantic (Pantic, 2001) propose un modèle asymétrique utilisant 4 paraboles ce qui porte le nombre

de paramètre à estimer à sept (xc , yc , θ , $w1$, $w2$, $h1$, $h2$). Le modèle complet, contour interne et contour externe, comprend onze paramètres.

Dans (Coianiz, 1996), le modèle proposé est similaire à celui de (Yuille, 1992) à la différence que les courbes sont des paraboles (figure 1.26). Cette solution permet une meilleure modélisation de la zone de l'arc de cupidon. Yokogawa et al. utilisent la même modélisation du contour supérieur et deux paraboles pour le contour inférieur (Yokogawa, 2007).

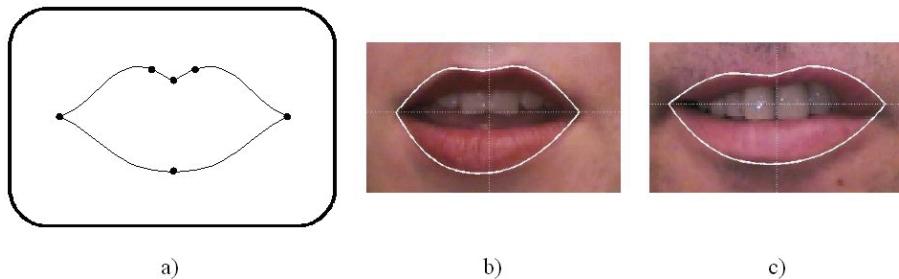


Figure 1.27 : Modèle paramétrique proposé par (Eveno, 2004)

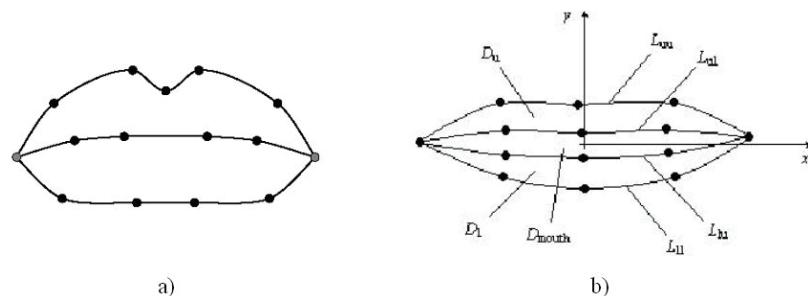


Figure 1.28 : Modèles paramétriques basés sur des splines, a) Modèle proposé par Vogt (Vogt, 1996), b) Modèle proposé par Malciu (Malciu, 2000).

Eveno (Eveno, 2004) a proposé un modèle composé de quatre courbes cubiques et d'une ligne brisée reliant cinq points clés du contour externe de la bouche (figure 1.27). Chaque courbe cubique est estimée par l'algorithme des moindres carrés à partir de la position de cinq points : une commissure, le centre (haut ou bas) du contour externe de la bouche, et trois points voisins de ce même centre. Ce modèle présente une bonne flexibilité et permet de décrire des variations importantes de la forme de la bouche. Vogt (Vogt, 1996) et Malciu (Malciu, 2000) utilisent, quant à eux, des splines pour modéliser les contours, haut et bas, de la bouche. Dans le cas de Vogt, sept points contrôlent le contour haut et six

le contour bas (figure 1.28-a). Malciu utilise cinq points de contrôle pour chaque contour (figure 1.28-b).

1.5.2.3 Modèles paramétriques : Initialisation

Une fois que le modèle paramétrique de bouche a été défini, l'étape suivante est l'initialisation du modèle. Pour les cas où il y a un modèle pour les bouches ouvertes et un modèle pour les bouches fermées, il faut tout d'abord identifier l'état de la bouche. L'algorithme doit ensuite initialiser la position du modèle dans l'image inconnue.

- Détection de l'état de la bouche :

Dans (Zhang, 1997), l'identification de l'état de la bouche résulte de l'étude d'une carte des contours de la zone de la bouche. Après avoir positionné les commissures de la bouche, une carte des contours est calculée à partir de la composante Y de l'espace $YCbCr$. Deux droites sont ensuite tracées, une droite joignant les commissures et une droite passant par le milieu de la bouche et perpendiculaire à celle joignant les commissures. L'intersection entre cette droite et les contours donnés par la carte donne le nombre de contours candidats de la bouche. Si le nombre est supérieur à deux, au dessus et au dessous de la ligne joignant les commissures, la bouche est considérée ouverte. Les candidats trouvés à l'étape précédente servent à initialiser les paraboles modélisant les contours internes (figure 1.21 et 1.22) et externes (figures 1.25).

Pantic (Pantic, 2001) et Coianiz (Coianiz, 1996) appliquent une transformation sur la teinte H pour déterminer la zone de la bouche et l'état de celle-ci. Les commissures sont positionnées en cherchant les jonctions des frontières, haute et basse, de la bouche à partir du gradient de la teinte. Ensuite, Pantic et Coianiz placent deux rectangles verticaux dans la zone de la bouche (figure 1.29-a) et font la somme le long des colonnes des valeurs dans la teinte transformée. Après avoir fait la moyenne d des deux sommes obtenues (figure 1.29-b et 1.29-c), si un seul maximum est trouvé, la bouche est considérée fermée (figure 1.29-b) sinon elle est considérée ouverte (figure 1.29-c). Ensuite, par seuillage sur d , les extrêmes des lèvres (deux dans le cas bouche fermée, quatre dans le cas bouche ouverte) servent à initialiser le contour interne (figure 1.21 et 1.22) et le contour externe (figure 1.26 pour

Coianiz et figure 1.25 pour Pantic). Yin détermine l'état de la bouche de manière analogue (Yin, 2002). Il calcule la somme le long des colonnes dans H sur un rectangle vertical placé au milieu de la zone de la bouche.

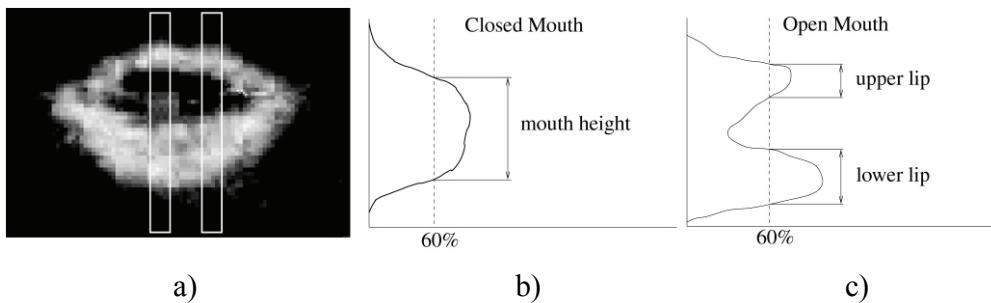


Figure 1.29 : Détection de l'état de la bouche (Pantic, 2001), a) Rectangles placés dans la zone de bouche, b) profil d de bouche fermée, c) profil d de bouche ouverte (Pantic, 2001)

Dans (Chen, 2004), l'état de la bouche est déterminé à partir d'étapes de seuillage effectuées sur les composantes *RGB*. L'auteur cherche à obtenir un masque des zones sombres de la région de la bouche. La zone interne de la bouche est censée être plus sombre que les lèvres. Ensuite, l'auteur trace trois droites, une droite verticale et deux droites diagonales, dans la région de la bouche. Ces droites donnent a priori 6 points d'intersection avec la zone sombre de la bouche. Si la distance est suffisamment grande entre ces points, la bouche est considérée comme ouverte.

Dans (Vogt, 1996), un réseau de neurones est entraîné sur les niveaux de teinte et de luminance d'images de référence afin de discriminer 5 classes de bouche : bouche fermée, bouche ouverte sans dents, bouche ouverte avec les dents occupant tout l'intérieur de la bouche, bouche ouverte avec les dents séparées par la cavité buccale, et bouche ouverte où seulement les dents supérieures sont visibles.

- Initialisation des paramètres des modèles paramétriques

L'initialisation est une étape critique pour l'optimisation des modèles paramétriques. Un modèle paramétrique doit au préalable être positionné approximativement avant de lancer la boucle d'optimisation. Dans la section 1.4, nous avons présenté des méthodes qui permettent de localiser la région de la bouche sur le visage par des approches « région ». Ce type de méthode est couramment utilisé pour positionner des modèles paramétriques.

Des auteurs ont proposé des méthodes différentes pour positionner leur modèle paramétrique. Dans (Coianiz, 1996), 6 points caractéristiques servent à positionner le modèle paramétrique de la figure 1.26. Deux points correspondent aux commissures, ils sont détectés par une analyse sur la chrominance. Les 4 autres points correspondent aux intersections entre la ligne verticale coupant la bouche en 2 et les contours externes et internes. Pour détecter ces points, Coianiz utilise la luminance. Dans (Pantic, 2001), l'auteur applique un algorithme de chaînage pour détecter grossièrement le contour externe de la bouche sur une grandeur colorimétrique dérivée de la teinte H . Une transformation est appliquée sur H pour faire ressortir les teintes rouges de l'image. Par la suite, l'auteur applique l'algorithme de chaînage dans cette nouvelle composante chromatique. Un germe est placé dans la partie basse de la zone de la bouche. Les commissures sont détectées quand la direction du contour obtenu par chaînage change. Les limites verticales de la bouche sont déterminées en projetant l'image de teinte sur la verticale. Zhiming et al. (Zhiming, 2002) utilisent les projections horizontales et verticales d'une composante chromatique, analogue à celle utilisée par (Pantic, 2001), pour positionner les commissures, et déterminer la boîte englobant la bouche (figure 1.30). Dans (Werda, 2007), les projections horizontales de la composante S servent à détecter les commissures.

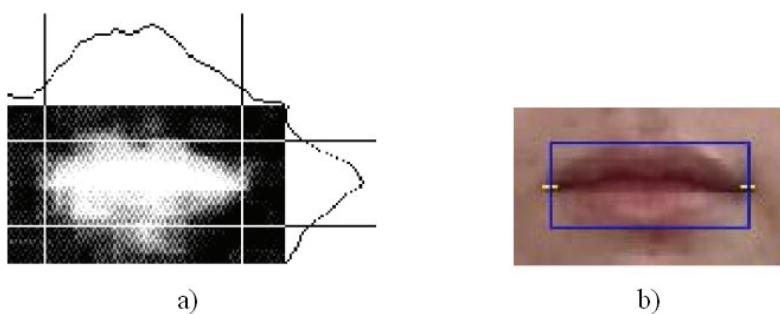


Figure 1.30 : Initialisation du modèle paramétrique (Zhiming, 2002), a) Projections horizontales et verticales de la composante chromatique, b) boîte englobant la bouche.

D'autres travaux proposent d'utiliser des snakes pour initialiser les modèles paramétriques. Les modèles paramétriques vont permettre de régulariser les contours modélisés par les snakes. Dans (Eveno, 2004), l'auteur définit un « jumping snake » pour extraire les 6 points nécessaires à l'estimation du modèle paramétrique modélisant le contour externe (figure 1.27). Le jumping snake est initialisé par un germe S^0 , qui peut être placé assez loin au dessus de la bouche (figure 1.31-a). Le snake subit alors une phase de croissance (figure

1.31-a). Ensuite un nouveau germe S^1 est déterminé. S^1 correspond au barycentre des points ajoutés lors de la croissance du snake. Ensuite, les opérations de croissance et de recherche d'un nouveau germe sont répétées jusqu'à ce que l'amplitude du saut entre 2 germes S^t et S^{t-1} soit inférieure à un certain seuil (figure 1.31-b). Le point bas du contour externe est ensuite trouvé en analysant le gradient le long de la ligne verticale passant par le centre du contour supérieur (figure 1.31-c).

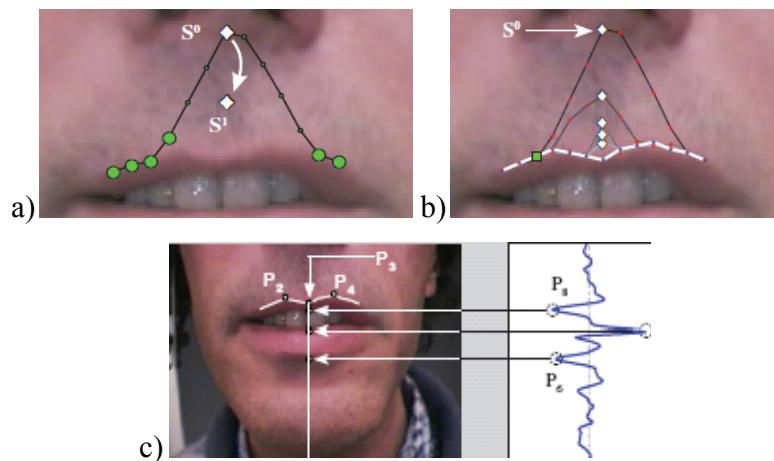


Figure 1.31 : Jumping snake (Eveno, 2004), a) Snake initial, b) Tracés des snakes pour les nouvelles positions du germe, c) Recherche du point bas du contour externe inférieur.

Dans (Jian, 2001) et (Salazar, 2007), les auteurs utilisent également un snake pour initialiser un modèle de contour externe de la bouche à 2 paraboles.

1.5.2.4 Modèles paramétriques : Optimisation

Comme dans le cas des snakes, les modèles paramétriques vont être optimisés itérativement par minimisation d'une fonction d'énergie interne et d'une fonction d'énergie externe. Dans la suite, nous nous proposons de donner des exemples de fonctions d'énergie utilisées dans les algorithmes de segmentation labiale par modèles paramétriques.

- Énergie interne

Dans le cas des snakes, les forces internes appliquent des contraintes locales dans le but de conserver une géométrie cohérente avec l'objet étudié pendant la convergence. Pour les modèles paramétriques, ces forces auront pour but de limiter les variations de forme. Dans

(Yuille, 1992), des termes d'énergie interne influent sur certaines propriétés du modèle. Ces propriétés sont la symétrie du contour supérieur, la position du centre de la bouche (au milieu des 2 commissures), l'épaisseur des lèvres (elle est considérée constante). Dans (Hennecke, 1994), une contrainte temporelle est ajoutée sur l'épaisseur des lèvres dans une séquence vidéo. Coianiz définit des contraintes sur les formes de bouche admissibles, le contour interne doit être strictement inclus dans le contour externe (Coianiz, 1996). Mirhosseini (Mirhosseini, 1997) et Vogt (Vogt, 1996) définissent une énergie potentielle, basée sur des mesures caractéristiques de la bouche, pour contrôler la forme du modèle. Dans (Malciu, 2000), un terme d'énergie, incluant des contraintes locales de symétrie (répartition uniforme des points de contrôle sur la courbe, figure 1.32-b) et d'élasticité (figure 1.32-a), est défini pour stabiliser les distances entre les points de contrôle du modèle.

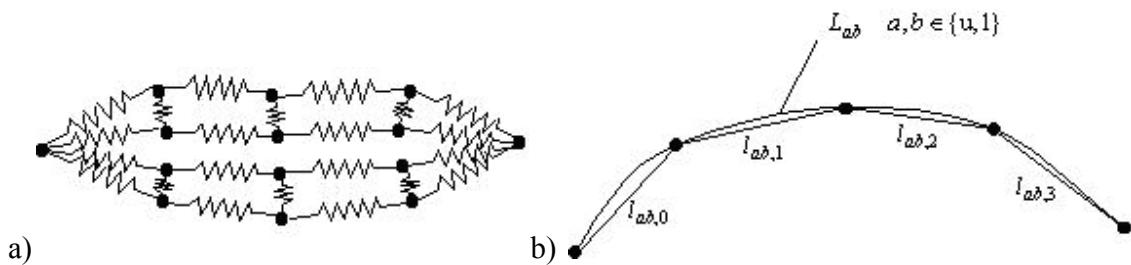


Figure 1.32 : Contraintes internes sur le modèle paramétrique (Malciu, 2000), a) Contraintes élastiques, b) contraintes de symétrie.

- Énergie externe

Dans la plupart des travaux sur la segmentation de la bouche, les contraintes géométriques intrinsèques imposées par le modèle sont suffisantes pour l'optimiser avec l'énergie externe seulement. L'expression de l'énergie externe permettra de faire converger le modèle paramétrique vers les caractéristiques saillantes de l'image, en particulier les contours de la bouche. Les forces externes qui servent à calculer l'énergie externe sont calculées à partir des données de l'image.

Comme pour les snakes, la force externe est couramment associée aux gradients de luminance. Dans (Yuille, 1992), la force externe est composée de 3 termes : une force dérivée du gradient de luminance, un champ de forces d'attraction des zones sombres et un champ de forces d'attraction des zones claires de l'image. Ces 3 champs combinés forment

la force externe. Dans (Mirhosseini, 1997), l'énergie externe est définie sur les zones sombres de l'image. Un terme est aussi défini sur l'intensité du gradient le long de la courbe du modèle. Dans (Malciu, 2000), un terme d'énergie prend en compte l'information de luminance globale de la bouche. Les zones claires correspondent aux lèvres, et les zones sombres correspondent à l'intérieur de la bouche. Un terme prenant en compte l'intensité du gradient le long du contour est également ajouté. Hennecke (Hennecke, 1994) utilise uniquement le gradient vertical pour calculer la force externe en tenant compte du signe de celui-ci pour optimiser le bon contour (externe haut, interne haut, interne bas et externe bas).

Bien que largement utilisé pour calculer la force externe, le gradient de luminance reste, par définition, très sensible aux variations d'éclairage. Pour s'affranchir de cette limitation, l'information de couleur apparaît pertinente quand elle est disponible. Eveno propose de combiner la pseudo-teinte \hat{H} et la luminance normalisée sur la zone de bouche, pour calculer un gradient particulier au contour haut de la bouche. Pour le contour externe inférieur, Eveno utilise uniquement le gradient de la pseudo-teinte (Eveno, 2004). Stillittano propose 2 gradients spécifiques aux contours internes (haut et bas) de la bouche (Stillittano, 2008).

Dans (Vogt, 1996), la force externe est donnée par le gradient calculé sur la carte de probabilité d'appartenance des pixels à la classe « lèvres » dans le cas d'une bouche ouverte. Sinon, le gradient d'intensité est utilisé. Dans (Yokogawa, 2007), une carte des contours de la bouche est calculée à partir d'un masque de la bouche obtenu par seuillage sur H . La force externe correspond alors à la différence entre la courbe donnée par le modèle et les contours donnés par la carte.

Comme nous l'avons vu, certains modèles paramétriques permettent de distinguer la peau, les lèvres, et l'intérieur de la bouche. Des critères, basés sur la couleur, ont été développés pour être minimisés lorsque les différentes régions de la bouche sont séparées. Dans (Coianiz, 1996) et (Pantic, 2001), trois zones sont définies, l'intérieur de la bouche, les lèvres et une région d'épaisseur constante entourant celles-ci. L'énergie est calculée sur l'information de chrominance des trois zones ; elle sera minimale lorsque les zones rouges correspondront aux lèvres. Dans (Zhang, 2001), deux fonctions d'énergie sont proposées, une pour l'état ouvert et une pour l'état fermé de la bouche. Elles sont définies par le

gradient de la luminance Y pondéré par la moyenne et la variance de Cr sur la zone considérée (lèvre haute, lèvre basse et intérieur de la bouche si celle-ci est ouverte). Yin construit un critère qui minimise les variances intraclasses des 3 régions définies par Zhang, et qui maximise les variances interclasses sur la composante H (Yin, 2002). Dans (Wu, 2002), le contour externe de la bouche est identifié après convergence d'un snake. Le contour interne est modélisé à l'aide d'un modèle paramétrique. La déformation du modèle est guidée par une fonction d'énergie basée sur les histogrammes des régions « lèvres » et « intérieur de la bouche ». Ces histogrammes sont obtenus par apprentissage sur une base d'images de bouche. Enfin, dans (Rao, 1995) et (Liew, 2000), l'optimisation est guidée par une fonction de probabilité qui est maximale lorsque la région des lèvres est entourée par le modèle paramétrique.

- Méthode d'optimisation

La méthode d'optimisation classique employée dans les algorithmes de segmentation labiale basés sur des modèles paramétriques est la descente du simplex (Yuille, 1992; Hennecke, 1994; Malciu, 2000). La descente du simplex est un algorithme d'optimisation qui ne nécessite que des évaluations de la fonction objective et non le calcul des dérivées. L'inconvénient de cette méthode est le coût en temps de calcul qui est très élevé, car elle nécessite un grand nombre d'évaluations de la fonction objective. Il faut également avoir à l'esprit que cette méthode ne garantit pas la convergence vers un minimum global de la fonction objective. Yuille, pour améliorer la robustesse, divise l'optimisation du modèle en 2 étapes (Yuille, 1992). En premier lieu, l'auteur recherche la position du centre de gravité, la largeur et l'orientation de la bouche. Ensuite, l'auteur applique l'algorithme de la descente du simplex aux paramètres restants.

Coianiz pratique une optimisation par descente du gradient utilisant la méthode des gradients conjugués (Coianiz, 1996). Les méthodes du type descente du gradient sont assez peu utilisées dans les problèmes d'optimisation de modèles paramétriques car elles nécessitent la connaissance du gradient de la fonction objective. La plupart du temps, ces dérivées ne sont pas disponibles.

D'autres auteurs effectuent, plus simplement, l'optimisation de leur modèle en recherchant les positions des points de contrôle qui minimisent la fonction d'énergie.

Dans (Vogt, 1996), une méthode sous-optimale est employée. Les points de contrôle du modèle sont positionnés aléatoirement à chaque itération. Si une nouvelle position donne une meilleure solution, alors celle-ci est conservée, sinon elle est rejetée.

Dans son initialisation, Zhang détecte des contours candidats des lèvres (Zhang, 1997). Il calcule ensuite les paraboles pour chacun des candidats et conserve celui qui minimise la fonction d'énergie adéquate (voir la sous-section précédente). Dans (Wu, 2002), la recherche du contour interne se fait par rapport au contour externe. Le contour interne est modélisé par 2 paraboles, 4 distances par rapport au contour externe servent à ajuster ces paraboles. Wu fait alors varier ces distances pour chercher la position optimale du contour interne par rapport à la fonction d'énergie citée dans la sous-section précédente. Dans (Eveno, 2004), l'optimisation du modèle paramétrique et la recherche des commissures de la bouche sont faites en même temps. Le modèle est calculé pour toutes les commissures candidates possibles à l'aide des points donnés par les contours, supérieur et inférieur, obtenus par l'algorithme du « jumping snake ». Les commissures candidates sont les points se trouvant sur la ligne de minimum de luminance passant entre les contours supérieur et inférieur. Le couple de commissures maximisant le flux du gradient à la courbe est conservé. Werda évalue toutes les positions possibles du modèle dans la boîte englobant la bouche et garde la réalisation qui maximise le flux du gradient à la courbe (Werda, 2007). Des auteurs comme Wark (Wark, 1998) et Stillittano (Stillittano, 2008) détectent des points clés pour calculer directement le modèle par moindres carrés. Stillittano calcule une carte binaire des contours de la bouche, et échantillonne ces contours pour estimer ses modèles de contours, externe et interne.

1.5.2.5 Discussion

Comme pour le cas des snakes, la principale limitation des modèles paramétriques est l'absence de garantie sur la convergence vers un minimum global. L'initialisation sera donc une étape critique pour la performance de l'extraction du contour des lèvres. Le choix du modèle aura également une incidence sur la précision de la modélisation de la bouche. Un compromis devra être trouvé entre complexité et vitesse de convergence. Plus le modèle sera simple, plus il convergera rapidement, mais au détriment de la précision.

Par rapport aux snakes, les modèles paramétriques présentent l'avantage d'intégrer une connaissance a priori sur la forme que l'on recherche. Après convergence, ce type de modèle offrira donc une solution cohérente. Dans le cas de la segmentation labiale, un modèle paramétrique sera par exemple intéressant en complément d'un snake.

1.6 Evaluation des algorithmes de segmentation de la bouche

Comme nous l'avons vu, il existe un très grand nombre d'algorithmes de segmentation de la bouche. La plupart du temps, les évaluations de ces algorithmes se limitent à des exemples visuels de convergence, et plus rarement à des comparaisons par rapport à des vérités-terrain sur des séries d'images. Il est donc souvent délicat d'apprécier la performance d'un algorithme par rapport à un autre. Nous considérons que l'évaluation est primordiale. Dans cette section, nous présenterons d'abord une série de bases d'images couramment utilisées en analyse faciale, puis, nous présenterons les différentes méthodes d'évaluation possibles d'un algorithme de segmentation des lèvres.

1.6.1 Bases d'images de visage

En l'état actuel de nos recherches, nous avons seulement trouvé deux bases d'images spécialement construites pour évaluer les algorithmes de segmentation des lèvres et de lecture labiale.

- La base CUAVE (Patterson, 2002) a été construite pour la reconnaissance de la parole avec les modalités visuelle et sonore. Cette base contient des séquences vidéo et audio de 36 sujets (17 femmes et 19 hommes). Des informations sur l'obtention de cette base sont disponibles sur la page internet : <http://ece.clemson.edu/speech>.
- La base LIUM (Daubias, 2003) a été construite en environnement non contrôlé. Elle contient des séquences avec des bouches maquillées et non maquillées. La base LIUM peut être obtenue gratuitement par les organisations académiques sur la page internet suivante : <http://www-lium.univ-lemans.fr/lium/avs-database/>.

On peut également citer plusieurs bases qui peuvent être utilisées pour l'évaluation des algorithmes de segmentation des lèvres :

-
- La base FERET (Philipps, 2000 ; Feret) contient un grande nombre d'images de visage dans différentes positions (de face, de côté) avec des conditions d'illumination et des expressions différentes. Cette base contient 1564 séries d'images représentant un total de 14126 images provenant de 1199 individus. Une fraction seulement de ces images peut être utilisée pour l'évaluation de la segmentation des lèvres. Cette base n'est pas disponible gratuitement. Des informations sont disponibles à la page internet :
<http://face.nist.gov/colorferet/>.
 - La base AR (Martinez, 1998) contient plus de 4000 images en couleurs correspondant à 126 individus (70 hommes et 56 femmes). Les sujets sont présentés de face avec des expressions différentes, différentes conditions d'illumination, et avec des accessoires cachant des parties du visage (lunettes de soleil, écharpe, ...). Cette base a été utilisée pour l'évaluation de la segmentation des lèvres par (Liew, 2003; Stillittano, 2008; Xin, 2005). La base AR est publiquement disponible et gratuite pour les chercheurs sur la page suivante :
http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html.
 - La base M2VTS (Pigeon, 1997) est composée d'images de 37 sujets différents. Ces images ont été prises à une semaine d'intervalle ou lors de changements physiques importants. Cette base a été utilisée pour l'évaluation de la segmentation des lèvres par (Lucey, 2000; Gordan, 2001; Pardas, 2001; Seguier, 2003; Wark, 1998). Cette base est publiquement disponible pour les applications non-commerciales. Pour toute information se référer à la page suivante :
<http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html>.
 - La base XM2VTS est une extension de la base M2VTS (Messer, 1999). Cette base a été utilisée pour l'évaluation de la segmentation des lèvres par (Kuo, 2005; Liew, 2003; Sadeghi, 2002). Cette base n'est pas gratuite. Pour toute information se référer à la page suivante :
<http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/>.
 - La base Cohn-Kanade AU d'expressions faciales codées (Kanade, 2000) est composée d'environ 500 séquences d'images provenant de 100 sujets et des données sur les expressions faciales. Cette base a été utilisée pour l'évaluation de la

segmentation des lèvres par (Pardas, 2001; Seyedarabi, 2006; Tian, 2000). Elle est disponible gratuitement à la page suivante :

http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html.

1.6.2 Evaluation des performances

L'évaluation et la comparaison des algorithmes de segmentation des lèvres sont des tâches complexes. Pour ce qui est de l'évaluation des performances, 3 familles de méthodes se dégagent : l'évaluation subjective, l'évaluation quantitative et l'évaluation globale par rapport à une application. Pour ce qui est des méthodes de comparaison, peu de protocoles existent. L'absence d'une véritable base de référence est certainement responsable de cette situation.

1.6.2.1 Evaluation subjective

La procédure d'évaluation classique des algorithmes de segmentation de la bouche est basée sur l'évaluation subjective du résultat par un expert humain. Par exemple, dans (Barnard, 2002; Liévin, 2004; Liew, 2003; Zhang, 2000), les auteurs affirment que leurs algorithmes ont été testés sur des ensembles d'images et que la segmentation a été correctement effectuée. Des images d'exemples illustrent alors la qualité de la segmentation.

Dans (Kuo, 2005), un système plus sophistiqué est présenté pour évaluer la performance de l'algorithme. Les résultats sont classés en cinq catégories, parfait, bon, moyen, mauvais et faux, suivant l'apparence générale du contour et la distance par rapport à 4 points clés (les coins de la boîte englobant la bouche). Sur la figure 1.33, on donne des exemples des résultats appartenant à ces catégories.

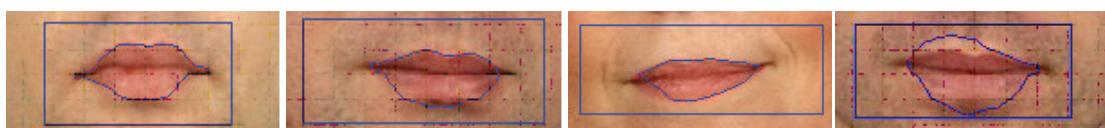


Figure 1.33 : Exemples d'évaluations subjectives (Kuo, 2005), De droite à gauche, parfait, bon, moyen et mauvais.

La limitation principale de ce genre de méthode d'évaluation est la subjectivité de l'expert. D'une personne à l'autre, l'évaluation pourra fortement varier. Le choix des frontières de

décision est dépendant de l'expert. Il faudra alors effectuer ce type d'évaluation sur un panel d'experts.

1.6.2.2 Evaluation quantitative

Pour une évaluation quantitative d'un algorithme de segmentation, on pourra considérer le contour dans son ensemble ou seulement des points clés, par exemple les commissures. Dans (Eveno, 2004), six points clés Q_i sont utilisés pour l'évaluation de la performance de l'algorithme d'extraction de contours (figure 1.34). Ces points clés ont été manuellement annotés par plusieurs opérateurs sur 300 images provenant de 11 sujets. Pour chaque point, la vérité-terrain est calculée comme la moyenne des positions cliquées par les opérateurs humains. Le résultat d'une segmentation est évalué en comparant la distance entre les points clés donnés par l'algorithme et les vérités-terrain. On donne ces erreurs pour le cas de (Eveno, 2004) dans la table 1.4. Ces erreurs sont normalisées par la largeur de la bouche.

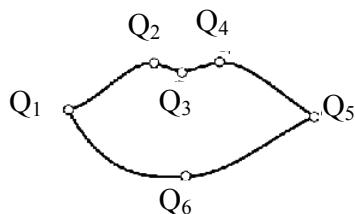


Figure 1.34 : Points clés Q_i utilisés pour l'évaluation (Eveno, 2004).

	Q ₁	Q ₂	Q ₃	Q ₄	Q ₅	Q ₆
Erreur (%)	4	2.4	1.8	1.9	3.3	3.6

Table 1.4 : Erreurs de détection sur les points clés (Eveno, 2004).

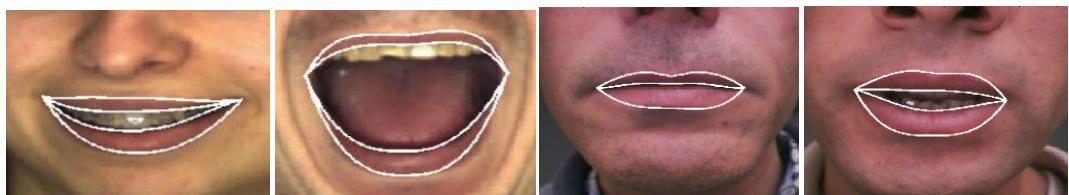


Figure 1.35 : Exemples de vérités-terrain (Stillitano, 2008). A gauche, on donne un exemple de sourire et de cri provenant de la base AR, à droite on donne des exemples d'images acquises avec un casque porté par le sujet et filmant la bouche.

La même méthode est employée dans (Yokogawa, 2007). La limitation soulignée par les auteurs eux-même, à l'utilisation de ce genre de méthode d'évaluation, vient de la possibilité d'erreurs lors de la segmentation manuelle des vérités-terrain.

Une évaluation quantitative peut aussi être réalisée sur l'ensemble du contour. Il faut alors disposer d'un contour de référence. Ce contour peut être obtenu par segmentation manuelle ou par un autre algorithme de segmentation. Quelle que soit la méthode, il faut prendre en compte le fait que l'on n'obtiendra jamais un unique contour de référence. Dans (Wu, 2002), le contour des lèvres est extrait manuellement. La qualité de la segmentation est donnée par le ratio du nombre d'erreurs sur le nombre de pixels dans le contour de référence. Un pixel est considéré comme erroné lorsqu'il n'est pas détecté à la fois dans les 2 contours. Wu (Wu, 2002) évalue son algorithme sur 126 images. Stillittano (Stillittano, 2008) utilise la même méthode sur un ensemble de 507 images de la base AR et sur 94 images provenant d'une base acquise avec un casque sur lequel est montée une caméra visant la bouche (figure 1.35). La limitation de cette approche est que pour un même nombre de pixels erronés, suivant la taille de la bouche, l'erreur sera alors plus ou moins grande. Une approche similaire est employée dans (Liew, 2003). Pour obtenir une vérité-terrain, les auteurs optimisent manuellement un modèle paramétrique sur 70 images tirées aléatoirement de leur base d'images. La performance est évaluée par 2 mesures : le pourcentage de recouvrement entre la région de référence et la région estimée et le ratio du nombre de pixels mal classés sur le nombre de pixels dans la région de référence. Wakasugi calcule le ratio $F_c = S_{diff}/L_c^2$ où S_{diff} et l'aire entre le contour externe de référence et le contour estimé, L_c est la longueur du contour de référence (Wakasugi, 2004).

1.6.2.3 Evaluation des performances par rapport à une application

Dans certains cas, la segmentation des lèvres est une étape dans un processus, par exemple pour la lecture sur les lèvres ou la reconnaissance d'émotions. La performance de la segmentation des lèvres est incluse dans la performance globale du système. Il faut alors prendre en compte le fait que la segmentation peut être « correcte » du point de vue de l'application visée, mais pas nécessairement fidèle. Dans (Brand, 2001), le but est d'évaluer le potentiel biométrique des lèvres pour la reconnaissance du sujet. L'algorithme est évalué sur le taux de reconnaissance des sujets. Dans (Chan, 1998; Chan, 2001; Chiou, 1997;

Nefian, 2002) le but est la reconnaissance de la parole. Le critère de performance considéré est alors le taux de reconnaissance des phonèmes considérés.

Gacon effectue une évaluation subjective en testant l'amélioration de l'intelligibilité d'une personne prononçant des numéros de téléphone par l'ajout d'un avatar à la bande sonore. Les mouvements des lèvres de l'avatar sont donnés par les contours segmentés par la méthode de Gacon. Le test est répété avec différents niveaux de bruit sur le son (Gacon, 2005). La même étude est réalisée avec le signal vidéo original pour quantifier la performance de la modélisation de la bouche.

Dans (Hsu, 2002), le but est de détecter des visages dans des images couleur et d'extraire des caractéristiques des visages (yeux, lèvres, ...) pour identification. L'algorithme est évalué sur le taux de détection des visages. Un visage est correctement détecté lorsque les ellipses englobant le visage et les yeux et la boîte englobant la bouche sont correctement trouvées.

Dans (Yin, 2002; Wu, 2002), la performance est évaluée sur des avatars. Le but est d'observer si les avatars ont un comportement réaliste. Pour cela, une comparaison entre la vidéo originale et celle avec l'avatar est réalisée subjectivement.

Wu compare les paramètres d'animation du visage générés après segmentation manuelle de la bouche avec ceux obtenus avec l'algorithme de segmentation (Wu, 2004).

Dans (Seyedarabi, 2006), les auteurs sont intéressés à reconnaître des *FAU* (Face Action Units, (Ekman, 1978)). Les lèvres segmentées sont utilisées pour calculer des caractéristiques géométriques. Ensuite, ces caractéristiques servent à construire un vecteur qui est utilisé pour classifier des *FAU* par un réseau de neurones. La performance de l'algorithme est évaluée sur le taux de reconnaissance des *FAU*.

1.7 Bilan

La modélisation des lèvres est un sujet toujours largement étudié. Pour ce qui est de la modélisation de la forme et de l'apparence de la bouche, de nombreux modèles ont été proposés sans qu'aucun ne devienne véritablement une référence. Concernant la segmentation des lèvres, nous avons présenté les deux grandes familles de méthodes utilisées : les approches « région » et les approches « contour ». D'après nos études, il ressort que les approches « contour » offrent une bonne précision mais que la robustesse de

ces méthodes n'est pas satisfaisante, car elle est trop dépendante de la précision de l'initialisation. Ces méthodes requièrent en outre le réglage d'un grand nombre de paramètres dépendant des conditions de l'environnement, du sujet et du système d'acquisition. Les méthodes dites « région », supervisées et non-supervisées, sont moins dépendantes de l'initialisation, mais elles n'offrent pas une modélisation fine des contours des lèvres.

Le problème de la modélisation et de la segmentation des lèvres est donc toujours ouvert. L'utilisation d'un seul type de méthode ou d'information n'est pas suffisant pour permettre une segmentation à la fois précise et robuste. En partant de ce constat, notre but dans cette thèse a été de proposer un ensemble de méthodes permettant une modélisation complète et robuste de la zone de la bouche, pouvant alors servir de base aux applications qui ont été présentées dans le chapitre d'introduction.

Chapitre 2. Segmentation région-contour de la bouche

2.1 Introduction

Dans le chapitre précédent, nous avons présenté les méthodes de modélisation et de segmentation de la bouche existantes. Nous avons pu voir que la majorité des travaux de modélisation de la bouche se concentrent sur le contour externe, bien que la zone interne de la bouche soit également importante, surtout dans l'optique de la lecture labiale. La première étape dans nos travaux sur la segmentation des contours des lèvres a été le développement d'une méthode de localisation de la bouche. Nous développerons dans la suite de ce chapitre notre méthode de localisation de la bouche et de segmentation d'un masque binaire des lèvres sur des images de visage en couleurs.

Dans la section 2.2, nous préciserons le cadre choisi pour notre étude. Nous détaillerons nos hypothèses de travail sur les images qui seront traitées par nos algorithmes.

À la section 2.3, nous proposerons une méthode pour augmenter le contraste entre les pixels de la peau et des lèvres. Nous reprendrons alors l'étude qui a été faite dans le chapitre 1 sur les espaces couleur.

Dans la section 2.4, nous aborderons le problème de la saillance des contours des lèvres. Nous proposerons d'employer un formalisme multi-échelle pour augmenter la robustesse de la modélisation des contours de la bouche.

Enfin, la section 2.5 présentera notre méthode de localisation et de segmentation des lèvres.

2.2 Segmentation labiale : Hypothèses de travail

Le but dans ce chapitre est d'effectuer la localisation et la segmentation de la bouche sur des images de visage. L'étude porte sur le cas d'images de visage en couleurs dans lesquelles la bouche est visible. Etant donné les applications possibles de la segmentation des lèvres, nous faisons l'hypothèse que les visages occupent la majorité de la surface de l'image soit que les visages ont été préalablement localisés et que la zone approximative de la bouche, la moitié basse du visage, est également connue.

La littérature est très abondante sur le problème de la détection des visages dans les images. On citera en particulier l'algorithme de Viola-Jones (Viola-Jones, 2001), implémenté dans OpenCV, qui est largement utilisé. Cette méthode est basée sur des vecteurs de caractéristiques, du type ondelettes de Haar, qui servent à entraîner un classifieur

AdaBoost. Les lecteurs intéressés par la détection de visage pourront également consulter le site <http://www.facedetection.com/> qui présente l'ensemble des méthodes de détection de visage dans des images avec un fond arbitraire.

En ce qui concerne la localisation grossière de la bouche, on considère qu'elle se trouve dans la moitié inférieure de la zone du visage. Aucune hypothèse n'est faite sur la taille ou l'échelle des images ni sur l'orientation du visage pourvu que la bouche soit visible. La figure 2.1 donne des exemples d'images d'entrée que nous avons étudiées.



Figure 2.1 : Exemples d'images de bouche utilisées dans notre étude

2.3 Grandeurs colorimétriques pour la modélisation des lèvres

Le choix des grandeurs colorimétriques est crucial pour tout algorithme de segmentation. Nous avons pu voir dans le chapitre 1, que beaucoup d'algorithmes de modélisation du contour externe de la bouche reposaient sur l'étude de la luminance. Le problème, lorsque l'on se base sur l'information de luminance, est la dépendance par rapport aux variations d'illumination de l'image. Suivant la direction de la source lumineuse par rapport au sujet, des réflexions spéculaires ou des ombres pourront apparaître sur les lèvres. Dans le cas d'une source de lumière située au dessus du sujet, on pourra voir apparaître des ombres sous la lèvre supérieure et sous la lèvre inférieure. Les réflexions et les ombres parasites rendront difficile l'identification de la zone de la bouche. La figure 2.2 présente les histogrammes normalisés de la luminance pour les ensembles de pixels de peau et des lèvres pour les mêmes images que celles utilisées à la section 1.2 du chapitre 1. On donne

également les variances intraclasses, interclasses ainsi que le ratio V_{intra}/V_{inter} dans la table 2.1. Si l'on compare les résultats de la table 2.1 et les histogrammes à ceux de l'étude effectuée dans le chapitre 1 sur les grandeurs colorimétriques dans le contexte de la séparation peau/lèvre, on peut conclure de ces résultats que la luminance n'est pas adaptée pour la segmentation des lèvres. Les tracés des histogrammes montrent un important recouvrement entre les distributions des pixels de la peau et des lèvres. Le rapport des variances V_{intra}/V_{inter} confirme cette impression. Dans le chapitre 1 nous avons vu que dans le meilleur des cas ce rapport était de 1,7 pour \hat{H} et dans le cas de la luminance ce rapport est de 38.5. On voit que, même sur un ensemble relativement faible d'images acquises avec la même source, les distributions se recouvrent presque entièrement. Cela indique d'importantes variations des niveaux de luminance pour les 2 ensembles de pixels. Un algorithme de localisation et de modélisation de la bouche basé sur la luminance sera a priori peu robuste aux variations des conditions de l'environnement. Pour localiser la bouche de manière robuste, il faudra disposer d'une ou de plusieurs grandeurs pour lesquelles le contraste peau/lèvre est fort et dans lesquelles les statistiques de ces ensembles sont stables.

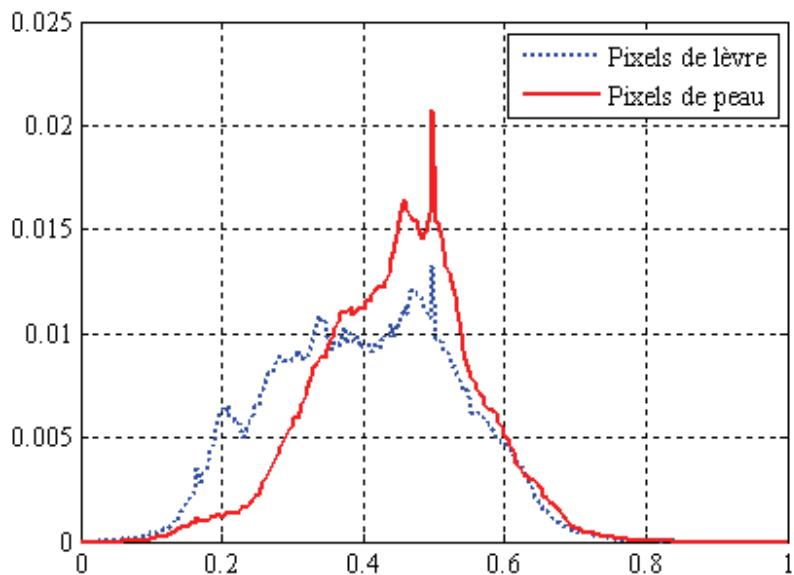


Figure 2.2 : Histogrammes de luminance des ensembles de pixels de la peau et des lèvres.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
Luminance	0.0128	$3.3172 \cdot 10^{-4}$	38.4922

Table 2.1 : Variances intraclasses, interclasses et V_{intra}/V_{inter} pour la luminance.

Dans la section 1.2 du premier chapitre, nous avons étudié les propriétés des espaces couleur classiques et des grandeurs colorimétriques destinées à augmenter le contraste entre la peau et les lèvres. Notre étude montre que les grandeurs *RGB* ne sont guère adaptées au problème de la séparation des lèvres et de la peau. Les valeurs des variances interclasses et intraclasses montrent que la séparation entre la peau et les lèvres est mauvaise. Le problème avec l'espace *RGB* vient du fait que les informations de teinte et de luminance sont mélangées. Par la suite, nous avons testé les propriétés de divers espaces et grandeurs colorimétriques sur la même base d'images dans le contexte de la séparation de la peau et des lèvres. Nous avons constaté que la pseudo-teinte \hat{H} et que la teinte \hat{U} offraient les meilleures performances pour séparer la peau et les lèvres. Par ailleurs, les rapports des variances intraclasses sur les variances interclasses indiquent que la séparation entre la peau et les lèvres est légèrement meilleure dans \hat{H} que dans \hat{U} . Les résultats présentés au chapitre 1 indiquent enfin que les propriétés des teintes de la peau et des lèvres sont relativement stables pour ces grandeurs colorimétriques. En particulier, les teintes rouges semblent plus fortes pour les lèvres que pour la peau. Dans \hat{U} , cela se traduit par une moyenne des pixels des lèvres inférieure à la moyenne des pixels de peau.

Si les grandeurs chromatiques \hat{H} et \hat{U} sont moins sensibles aux variations de luminance, leur calcul repose sur les grandeurs *R*, *G*, *B* dont on sait qu'elles sont corrélées avec la luminance.

L'algorithme allongement-décorrélation (decorrelation stretch, (Gillespie, 1986)) permet d'accentuer les différences entre les canaux des images produites par des systèmes multispectraux en éliminant la corrélation entre les différents canaux. Dans notre cas, nous disposons d'images de bouche dans l'espace *RGB*. Pour chaque pixel de l'image d'entrée, nous disposons d'un vecteur composé de 3 niveaux dans *RGB*. L'algorithme allongement-décorrélation permet de trouver une transformation de l'espace *RGB* vers un espace dans lequel la corrélation entre les composantes a été supprimée. Une fois les pixels projetés dans cet espace, les composantes sont normalisées par leurs variances. Enfin, on applique la

transformation inverse pour ramener les vecteurs normalisés dans l'espace de départ. La transformation inverse, appliquée aux vecteurs normalisés, permet de préserver la cohérence des teintes de l'image, tout en maximisant les différences de couleur.

Dans un premier temps, il nous faut déterminer la matrice de corrélation entre les canaux *RGB* à partir des pixels de l'image d'entrée. Soit N le nombre de pixels et $L=3$, le nombre de composantes couleur de l'image d'entrée. On calcule d'abord la matrice de covariance *COV* des vecteurs des données d'entrée. Les éléments de la matrice sont les coefficients COV_{ij} . La matrice de corrélation *CORR* s'exprime alors de la manière suivante :

$$\left\{ \begin{array}{l} CORR = \begin{bmatrix} CORR_{1,1} & \cdots & CORR_{1,3} \\ \vdots & \ddots & \vdots \\ CORR_{3,1} & \cdots & CORR_{3,3} \end{bmatrix} \\ CORR_{i,j} = \frac{COV_{i,j}}{(COV_{i,i} * COV_{j,j})^{1/2}} \end{array} \right. \quad (15)$$

On calcule ensuite les valeurs propres et les vecteurs propres de la matrice de corrélation. La matrice *ROT*, composée des vecteurs propres de la matrice de corrélation, permet alors de projeter les pixels de l'image d'entrée sur un nouvel espace où les composantes sont décorrélées. Soit le vecteur de normalisation Ω , il est composé des inverses des racines carrées des valeurs propres de la matrice de corrélation. En pratique, cela correspond aux écarts-types des données projetées sur les vecteurs propres de *CORR*. Au besoin les éléments de Ω peuvent être multipliés par des coefficients pour obtenir des écarts types particuliers. La transformation finale *T* s'écrit alors :

$$T = ROT^T \Omega ROT \quad (16)$$

La transformation *T* est d'abord appliquée au vecteur composé des niveaux moyens dans les trois composantes *R*, *G* et *B*. Le résultat permet de déterminer les décalages nécessaires pour recadrer les composantes *RGB* décorrélées entre les niveaux 0 et 255. *T* est, par la suite, appliquée à tous les pixels de l'image.

Nous avons répété l'étude présentée au chapitre 1 sur notre base d'images, mais avec, cette fois, les composantes *RGB* décorrélées R_{decorr} , G_{decorr} , B_{decorr} . Ensuite, nous avons recalculé les variances intraclasses et interclasses ainsi que les rapports V_{intra}/V_{inter} pour les

grandeurs colorimétriques spécifiques, étudiées au chapitre 1. Ces grandeurs ont été calculées à partir de R_{decorr} , G_{decorr} , B_{decorr} normalisées, au préalable, entre 0 et 1. Les résultats sont donnés dans la table 2.2. La figure 2.3 présente les tracés des histogrammes des distributions des pixels de peau et des lèvres pour ces nouvelles grandeurs colorimétriques.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
R_{decorr}	$9.46 \cdot 10^{-3}$	$2.38 \cdot 10^{-3}$	3.97
G_{decorr}	0.0222	0.0313	0.7095
B_{decorr}	$1.18 \cdot 10^{-2}$	$1.41 \cdot 10^{-3}$	8.38
Cb_{decorr}	0.0087	0.0050	1.7536
Cr_{decorr}	0.0078	0.0066	1.1787
H_{decorr}	0.0079	0.0016	4.9481
\hat{H}_{decorr}	0.0078	0.0114	0.6866
\hat{U}_{decorr}	0.0331	0.0529	0.6256

Table 2.2 : Variance intraclasses, variance interclasses et V_{intra}/V_{inter} pour les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , \hat{H}_{decorr} et \hat{U}_{decorr} .

D'une manière générale, les résultats de la table 2.2 montrent une nette diminution du rapport V_{intra}/V_{inter} pour toutes les grandeurs colorimétriques. Il y a bien une augmentation du contraste entre la peau et les lèvres lorsqu'on applique l'algorithme allongement-décorrélation. On constate, sur les grandeurs R_{decorr} , G_{decorr} , B_{decorr} obtenues après transformation, que c'est la grandeur G_{decorr} qui offre le plus fort contraste peau/lèvres. Pour la teinte H_{decorr} , issue de la transformation vers l'espace HSV , on note une légère dégradation du contraste par rapport à H . Ce phénomène s'explique par une augmentation du bruit sur les grandeurs R_{decorr} , G_{decorr} , B_{decorr} . La transformation vers l'espace HSV est non-linéaire, ce qui la rend sensible au bruit. Le bruit sur H_{decorr} sera donc d'autant plus important lorsqu'on appliquera la transformation vers l'espace HSV à partir des grandeurs R_{decorr} , G_{decorr} , B_{decorr} . Pour les composantes chromatiques proposées spécifiquement pour la segmentation des lèvres, on remarque un léger gain par rapport à la composante G_{decorr} . Pour illustrer les résultats de la table 2.2, la figure 2.3 présente les tracés des histogrammes des distributions des pixels de la peau et des lèvres des grandeurs décorrélées. Pour les besoins des tracés, les différentes composantes ont été normalisées entre 0 et 1.

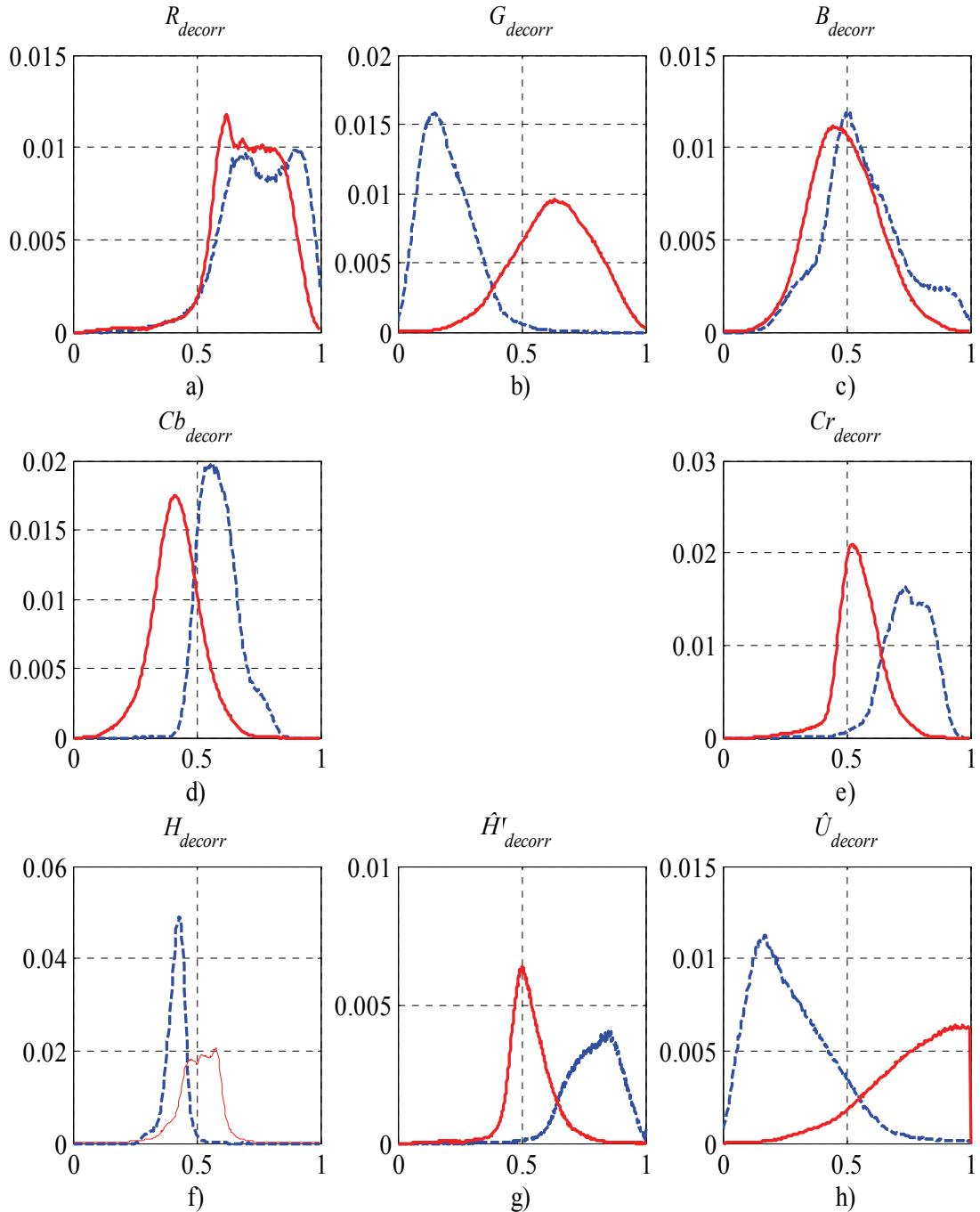


Figure 2.3 : Tracés des histogrammes des distributions des pixels de la peau (en rouge) et des pixels des lèvres (en bleu) pour les grandeurs chromatiques obtenues à partir des grandeurs R_{decorr} , G_{decorr} , B_{decorr} , a) R_{decorr} , b) G_{decorr} , c) B_{decorr} , d) Cb_{decorr} , e) Cr_{decorr} , f) H_{decorr} , g) \hat{H}_{decorr} et h) \hat{U}_{decorr} .

Les objectifs visés dans ce chapitre sont, d'une part, de localiser la bouche et, d'autre part, de segmenter un masque binaire des lèvres. Dans la suite de notre étude, nous avons choisi de travailler avec la grandeur \hat{U}_{decorr} , calculée à partir des grandeurs *RGB* décorrélées, pour localiser la bouche et segmenter un masque binaire de la bouche. Avec cette grandeur, nous avons observé une bonne stabilité des statistiques des ensembles de pixels de la peau et des lèvres sur un ensemble de sujets. De plus, cette grandeur offre un contraste important entre la peau et les lèvres. Nos ensembles de pixels de peau et des lèvres ont été composés d'échantillons provenant de 20 sujets. Nous pouvons conclure que cette grandeur est robuste aux changements de sujets et aux variations des conditions de l'environnement. Le choix de \hat{U}_{decorr} nous semble le meilleur, pour localiser et segmenter un masque des lèvres.

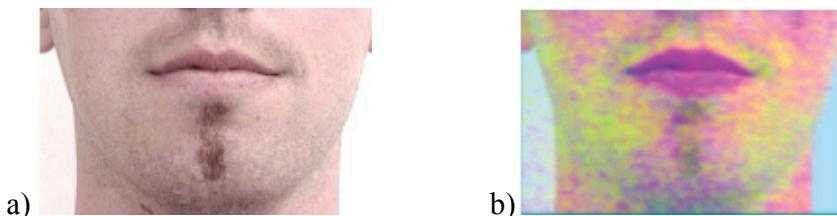


Figure 2.4 : Exemple d'une image de bouche dans *RGB*, a) image sans traitement, b) image après décorrélation et normalisation.

La figure 2.4 présente un exemple d'image de bouche dans *RGB* sans traitement et après décorrélation et normalisation des composantes couleur. On constate l'augmentation très importante du contraste entre la peau et les lèvres. On note également une augmentation du niveau du bruit sur l'image. Dans la suite de ce rapport, nous considérons que les grandeurs colorimétriques sont toutes issues des composantes *RGB* décorrélées et normalisées, \hat{U} correspondra alors à \hat{U}_{decorr} .

2.4 Gradients Multi-échelle pour la modélisation des contours de la bouche

Dans la section 1.3 du chapitre 1, nous avons présenté plusieurs gradients développés pour modéliser les contours de la bouche. Un des problèmes soulevés dans le chapitre 1, relatif à la modélisation des contours de la bouche, est que, bien souvent, les contours des lèvres sont peu marqués. C'est-à-dire que la transition peau/lèvre est très douce. Ceci se traduit par

des gradients de faible intensité. L'identification et la modélisation des contours de la bouche seront alors difficiles. Pour illustrer cette remarque, la figure 2.5 présente le cas d'une image de bouche ainsi que le gradient de \hat{U} .

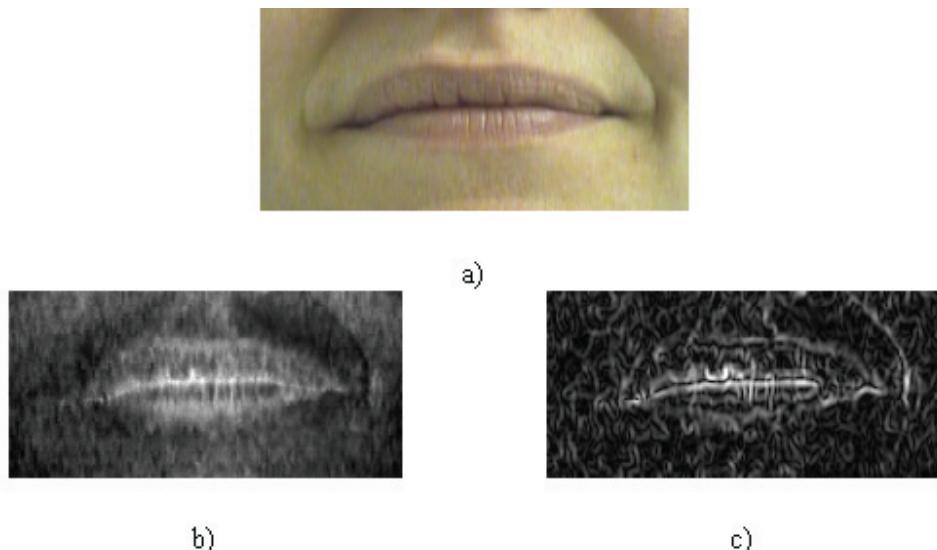


Figure 2.5 : Exemple de calcul du gradient de \hat{U} pour une image de bouche : a) Image de bouche, b) Teinte \hat{U} , c) Intensité du gradient de \hat{U} .

On voit, sur l'image de la figure 2.5-c, que l'intensité du gradient est faible sur le contour externe de la bouche et que le niveau du bruit est important. La modélisation du contour externe sera donc très difficile quelle que soit la méthode employée.

Pour aborder ce problème, il est possible de considérer que, pour le cas de la figure 2.5, l'échelle à laquelle on observe le phénomène d'intérêt, la transition peau-lèvre, n'est pas pertinente. L'idée serait alors de se placer à une échelle plus pertinente, de manière à mieux observer le phénomène. En filtrant l'image de départ, qui correspond à la plus grande échelle (niveau de détail maximal), par un ensemble de filtres gaussiens, dont la variance est croissante, on obtient alors une famille d'images (Scale-Space Representation (Lindeberg, 1998)) dont le niveau de détail est diminué progressivement. Avec cette représentation, tout se passe comme si l'échelle à laquelle on observe le phénomène était diminuée progressivement. Pour une image donnée $f(x,y)$, la représentation multi-échelle (Scale-Space Representation) L de l'image est définie de la manière suivante :

$$L(x, y, t) = g(x, y, t) * f(x, y) \quad (17)$$

où $*$ est l'opérateur de convolution et $g(x, y, t)$ est un filtre gaussien de la forme suivante:

$$g(x, y, t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2+y^2}{2t}} \quad (18)$$

où $t=\sigma^2$ est le paramètre d'échelle, σ est la variance de la gaussienne et $L(x, y, 0)=f(x, y)$. A partir de cette représentation, on peut alors exprimer les dérivées multi-échelles pour n'importe quelle échelle $t=\sigma^2$ de la manière suivante :

$$L_{x^\alpha y^\beta}(x, y, t) = \partial_{x^\alpha y^\beta} L(x, y, t) = (\partial_{x^\alpha y^\beta} g(x, y, t)) * f(x, y) \quad (19)$$

où α et β correspondent respectivement aux ordres des dérivées selon x et y . Dans notre cas, on s'intéresse au gradient de $L(x, y, t)$, c'est-à-dire aux dérivées $L_x(x, y, t)$ et $L_y(x, y, t)$ et à l'énergie du gradient $|\nabla L(x, y, t)|^2 = L_x^2(x, y, t) + L_y^2(x, y, t)$. À la figure 2.6, on présente les images d'intensité du gradient de $\hat{U}(x, y)$ de la figure 2.5-b. Les images d'intensité ont été normalisées entre $[0, 1]$ et σ varie de 1 à 10. La normalisation de l'intensité n'a pour but que d'améliorer la visualisation des images d'intensité. La question de l'amplitude sera traitée dans la suite de cette section. On constate sur la figure 2.6 que, les contours de la bouche deviennent plus saillants lorsque σ augmente. Dans le même temps, les détails de la forme deviennent de moins en moins précis. Dans ce chapitre, nous sommes intéressés par la localisation et la segmentation des lèvres. Nous chercherons donc à maximiser l'intensité des gradients sur les contours de la bouche.

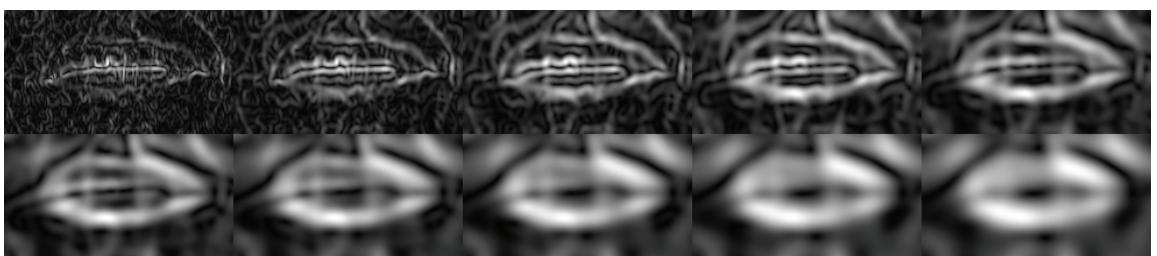


Figure 2.6 : Calcul du gradient de \hat{U} , pour σ ($t=\sigma^2$) allant de 1 à 10.

La figure 2.7 présente une coupe verticale de l'image d'intensité $|\nabla L(x, y, t)| = \sqrt{L_x^2(x, y, t) + L_y^2(x, y, t)}$ du gradient de \hat{U} de l'image de bouche pour des valeurs de $t=\sigma^2$ croissantes (de $\sigma = [1, \dots, 10]$). On observe que l'intensité des maximums locaux correspondant aux contours externes supérieur et inférieur de la bouche diminue quand σ augmente. Cependant, sur la figure 2.6, nous avons constaté que la saillance du contour externe par rapport au reste de l'image s'améliore lorsque σ augmente, bien que les détails de la forme soient de moins en moins précis.

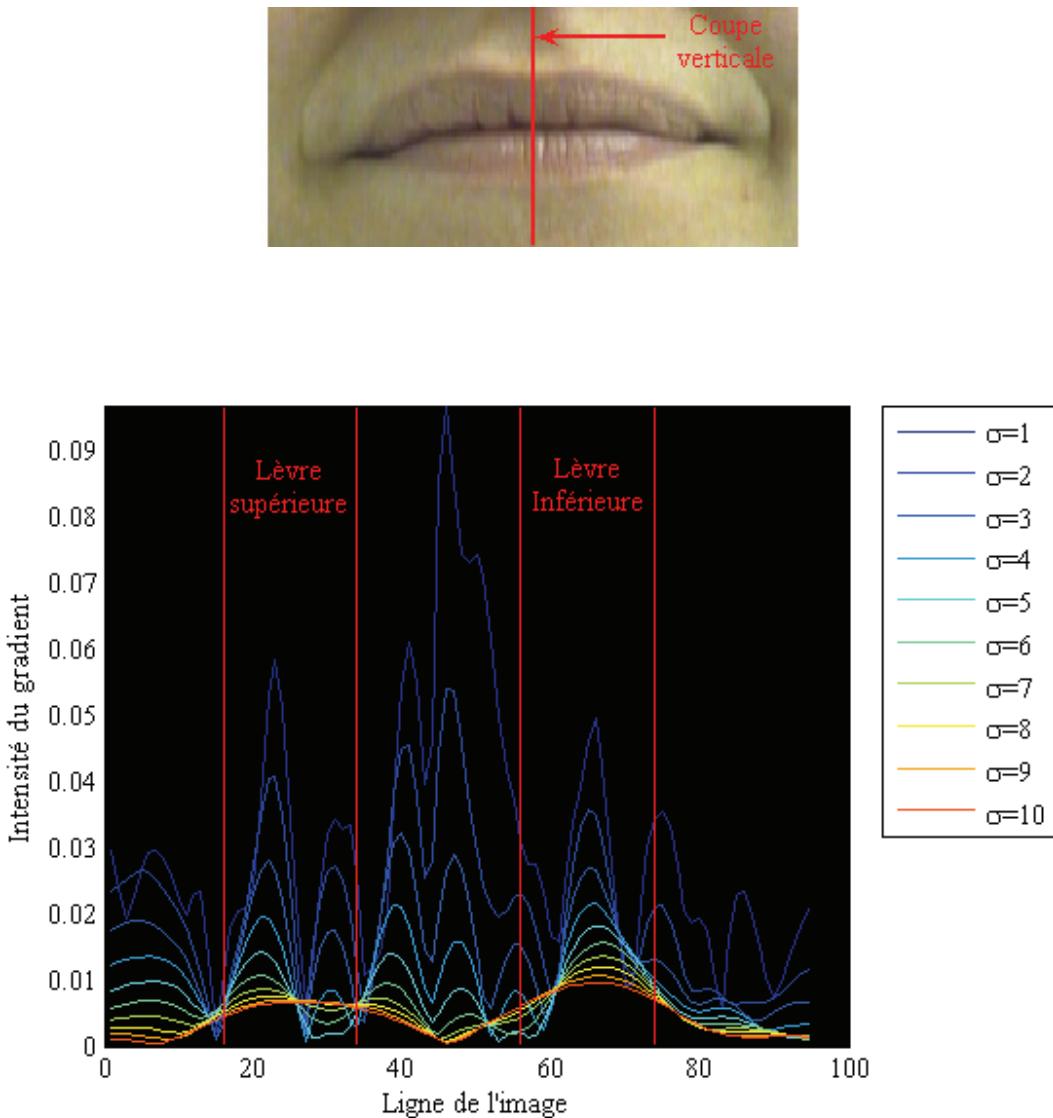


Figure 2.7 : Intensité des gradients $\nabla L(x, y, t)$ de \hat{U} le long de la coupe centrale de l'image de bouche pour des valeurs de σ allant de 1 à 10.

Une pondération entre les gradients est nécessaire pour que les échelles où l'information pertinente est présente soient prépondérantes dans la caractérisation, par exemple, du contour externe de la bouche. On voit que, sans normalisation, l'intensité des gradients va décroître lorsque σ va augmenter. Les grandes échelles seront toujours prépondérantes, même si les contours de la bouche sont très peu saillants.

Lindeberg (Lindeberg, 1998) propose une méthodologie générale pour choisir les échelles intéressantes lors de l'extraction de caractéristiques dans les images (contours, coins, blobs). Le principe est d'étudier les dérivées γ -normalisées, exprimées par l'opérateur suivant :

$$\partial_{\xi,\gamma-norm} = t^{\gamma/2} \partial_x \quad (20)$$

où γ est un paramètre à déterminer suivant le problème et $\xi = x/(t^{\gamma/2})$ et un changement de variable. La sélection de l'échelle d'intérêt revient alors à rechercher des extrema locaux des dérivés γ -normalisées par rapport au paramètre d'échelle t .

Pour justifier l'effet de cette pondération des dérivées, on se propose d'observer l'effet d'une transformation d'échelle $x' = sx$, avec s facteur d'échelle tel que $f'(sx) = f(x)$, sur les expressions des dérivées γ -normalisées. On a alors $t' = s^2 t$. Les expressions des dérivées γ -normalisées de $f'(x', y')$ sont de la forme suivante :

$$\partial_{\xi^m} L(x, t) = s^{m(1-\gamma)} \partial_{\xi'^m} L'(x', t') \quad (21)$$

Cette expression montre que les extrema locaux des dérivées γ -normalisées, par rapport à l'échelle, seront préservés à un facteur de pondération près. Il est alors possible de construire un détecteur invariant aux changements d'échelle pour extraire des caractéristiques particulières (contours, coin, ...) à une échelle donnée, en utilisant les dérivées γ -normalisées (Lindeberg, 1998). Le paramètre γ sera réglé en fonction de la caractéristique étudiée. D'une manière générale, on fait l'hypothèse que les dérivées γ -normalisées, à une transformation d'échelle s près, sont identiques. Ceci implique que γ est fixé à 1.

Pour le cas particulier de la modélisation des lèvres nous sommes intéressés par les expressions des gradients γ -normalisés $\nabla_{norm}L(x,y,t) = L_{x,\gamma-norm}(x,y,t) + L_{y,\gamma-norm}(x,y,t)$ avec $L_{x,\gamma-norm}(x,y,t) = t^{\gamma/2} \cdot L_x(x,y,t)$ et $L_{y,\gamma-norm}(x,y,t) = t^{\gamma/2} \cdot L_y(x,y,t)$, $f(x,y) = \hat{U}(x,y)$ et $\gamma=1$. Sur la figure 2.8, nous avons tracé les courbes d'intensité des gradients γ -normalisés de $\hat{U}(x,y)$ sur la coupe verticale centrale de l'image de bouche de la figure 2.5-a pour les mêmes valeurs de $t=\sigma^2$ croissantes (de $\sigma=[1,\dots,10]$).

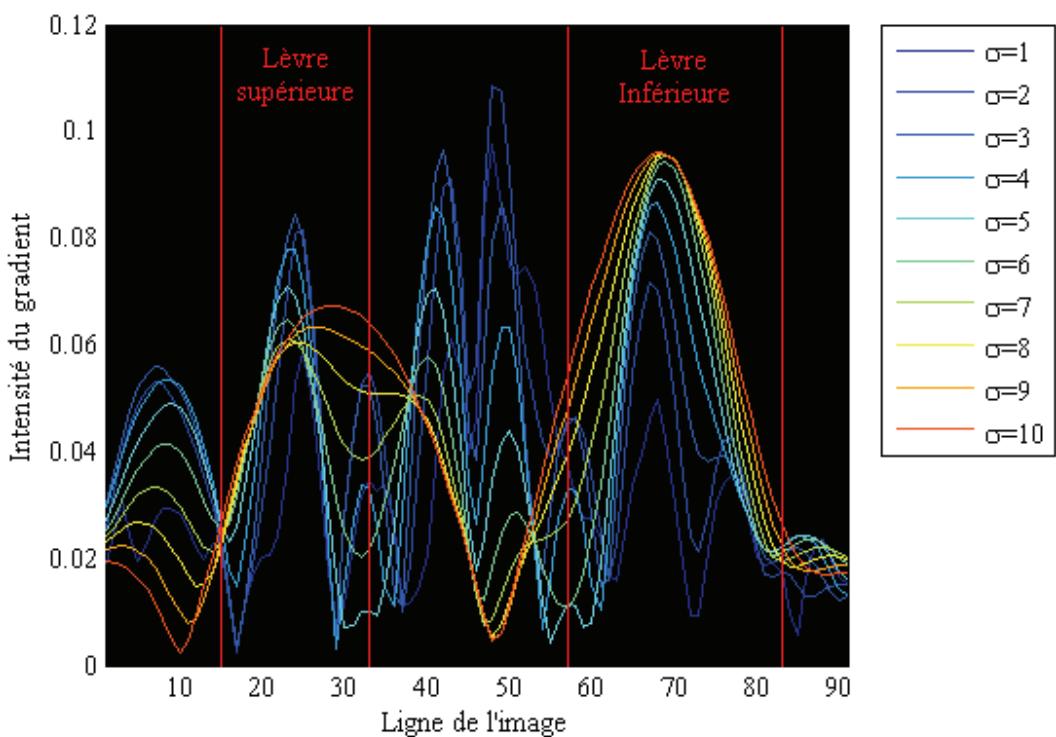
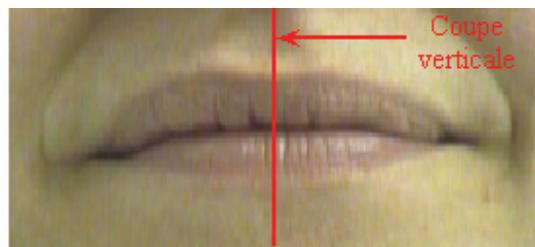


Figure 2.8 : Intensité des gradients $\nabla_{norm}L(x,y,t)$ de \hat{U} le long de la coupe centrale de l'image de bouche pour des valeurs de σ allant de 1 à 10.

Pour le cas du contour externe de la lèvre supérieure, on remarque que l'intensité maximale du gradient est atteinte pour $\sigma=3$. Pour le contour externe inférieur, l'intensité est maximale pour $\sigma=10$. Pour le contour externe supérieur, on constate sur la figure 2.6 que $\sigma=3$ correspond bien à une échelle où le contour est bien défini et détaillé. Pour le contour externe inférieur, $\sigma=10$ correspond à une petite échelle où la lèvre inférieure est très saillante mais pour laquelle le contour est très flou. On constate aussi que les positions spatiales des maximums locaux sont stables pour les échelles où l'intensité est la plus importante.

Pour modéliser globalement les contours des lèvres sans a priori, il sera difficile de privilégier une seule échelle. L'exemple étudié dans cette section montre que les échelles d'intérêt peuvent être différentes pour le contour externe supérieur et le contour externe inférieur et que plusieurs échelles peuvent être intéressantes. D'une manière générale, en utilisant une famille de gradients γ -normalisés, en partant des grandes échelles avec le maximum de détails et en incluant les petites échelles pour lesquelles l'intensité est importante, nous serons capables de décrire fidèlement les contours de la bouche. Les échelles les plus grandes, pour lesquelles l'intensité est plus faible mais avec le plus de détails, permettront une modélisation locale fine. Les petites échelles, modélisant grossièrement les contours mais avec de fortes intensités sur les contours, exercent quant à elles une force d'attraction importante sur des modèles de contour. La difficulté sera de choisir le nombre d'échelles à prendre en compte.

Par exemple, dans le cas du contour externe inférieur de la bouche, l'échelle $\sigma=10$ permet de localiser grossièrement le contour. En effet, on constate que la largeur ainsi que l'intensité de la transition sont maximales. À cette échelle, le gradient exercera une force d'attraction importante sur un modèle de contour externe inférieur. Les échelles plus grandes, dans lesquelles le niveau de détails est plus important, permettront d'affiner le détail du contour lors de la convergence. Dans la suite du chapitre, nous traiterons le problème de la localisation et de la segmentation des lèvres sur les images de visage. Une section sera consacrée au choix du nombre d'échelles à prendre en compte dans la partie consacrée à la segmentation région-contour des lèvres.

En conclusion, on voit, au travers de l'exemple étudié dans cette sous-section, que pour modéliser la bouche, une seule échelle n'est pas suffisante et que l'échelle de départ de

l'image n'est pas toujours adaptée à l'indice visuel recherché. Pour modéliser correctement les contours des lèvres sans a priori sur l'échelle, la solution proposée est de s'intéresser à des familles de gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ à différentes échelles.

2.5 Segmentation région-contour des lèvres

2.5.1 Méthodologie

Dans cette étude, nous faisons l'hypothèse que la moitié inférieure du visage a été localisée préalablement, cf. la section 2.1 de ce chapitre. Le premier problème à résoudre est la localisation de la bouche. Nous avons vu que les approches « contour » sont sensibles à l'initialisation et que la robustesse est dépendante de la distance entre l'initialisation et le résultat recherché. De plus, les propriétés de la teinte \hat{U} pour séparer la peau des lèvres, à savoir une moyenne de la distribution de teinte des pixels des lèvres inférieure à celle de la peau et une bonne séparation entre les 2 classes de pixels, permettent d'envisager une approche région non-supervisée pour localiser la zone de la bouche. Une approche supervisée ne nous semble pas pertinente dans le cas de la localisation de la bouche. Une telle approche nécessiterait l'entraînement d'un modèle de teinte sur une base de pixels. Yang a montré que le rendu des couleurs était très dépendant des dispositifs d'acquisition (Yang, 1996). Une étape de normalisation sera donc nécessaire, dans le cas d'une image inconnue, pour pouvoir appliquer le modèle de manière robuste. Une étape permettant d'identifier globalement le visage et les lèvres est donc de toute façon nécessaire.

Dans notre cas, nous avons choisi une approche combinée région-contour basée sur la teinte \hat{U} , dont nous avons pu voir qu'elle permettait de séparer au mieux les ensembles de pixels de la peau et des lèvres. Nous avons développé une approche hiérarchique de type descendante exploitant la teinte \hat{U} en 2 étapes (figure 2.9). Une première classification non-supervisée basée sur un modèle de mixture de gaussiennes de la distribution de teinte de l'image d'entrée permet de déterminer la zone de visage. La deuxième étape consiste à extraire un masque des lèvres de manière itérative, à partir de la zone de visage. Cette deuxième étape est également basée sur une classification non-supervisée exploitant les propriétés de \hat{U} et les informations fournies par les gradients γ -normalisés. Le principe est de localiser la zone de la bouche, d'extraire le masque identifiant les lèvres, d'extraire le

contour du masque et d'en optimiser la forme à l'aide des gradients γ -normalisés. Enfin, une étape supplémentaire permet d'affiner la précision du masque.

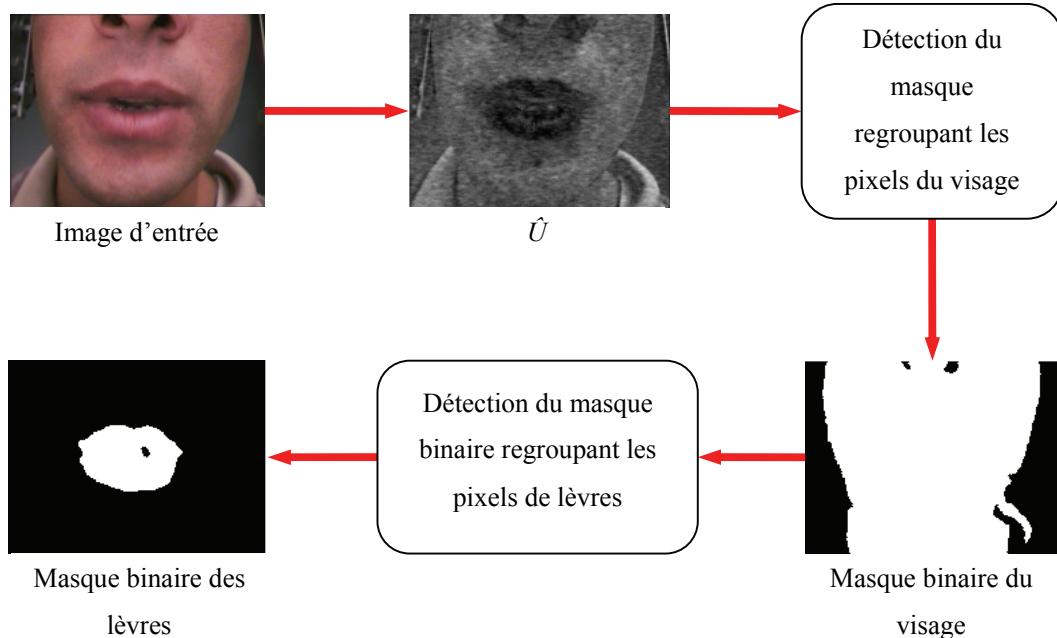


Figure 2.9 : Schéma bloc résumant les étapes de la segmentation des lèvres.

2.5.2 Localisation de la zone du visage

Dans la section 2.1, nous avons fait l'hypothèse que les visages ont été préalablement détectés et que l'on se concentre sur la moitié inférieure du visage. Dans l'image, peuvent être présents, des pixels du visage et de l'arrière plan. Nous avons également vu que la teinte \hat{U} permettait de bien séparer les pixels des lèvres et les pixels de la peau. Pour pouvoir exploiter cette propriété, il est important de réduire la recherche de la zone de la bouche à des zones identifiées comme « visage », pour supprimer l'influence de l'arrière plan. L'hypothèse qui est faite sur le cadrage des images implique que la majorité des pixels de l'image appartiennent à la zone du visage. On suppose que les pixels de la teinte \hat{U} de l'image d'entrée peuvent être séparés en M classes. On se retrouve alors face à un problème de classification avec M classes. Étant donné que les pixels du visage sont

considérés majoritaires dans l'image, une approche statistique semble appropriée pour modéliser la zone de visage.

2.5.2.1 Classification basée sur un mélange de gaussiennes

Soit une image I quelconque. I est composée des pixels $P = [\hat{u}_1 \dots \hat{u}_N]$ appartenant à \mathbb{R}^d , d dimension des vecteurs représentant les pixels. Si on fait l'hypothèse que l'ensemble P des pixels peut être séparé en M classes, et que la distribution des pixels de l'image peut être modélisée par des gaussiennes G_i , de moyenne μ_i et d'écart type σ_i avec $i = [1 \dots M]$, alors, pour l'ensemble des N pixels de l'image, la distribution $f(\hat{u})$ pourra être modélisée par un mélange pondéré des gaussiennes $G_i(\hat{u})$:

$$f(\hat{u}) = \sum_{i=1}^M w_i G_i(\hat{u}) \quad (22)$$

où w_i sont les poids du mélange, tels que $w_i > 0$ et $\sum_{i=1}^M w_i = 1$. Ces poids représentent les proportions des différentes classes et G_i est la distribution gaussienne modélisant la classe i :

$$G_i(\hat{u}) = \frac{1}{(2\pi)^{d/2} \times \det(\Sigma_i)^{1/2}} \exp\left(-\frac{1}{2} (\hat{u} - \mu_i)^T \Sigma_i^{-1} (\hat{u} - \mu_i)\right) \quad (23)$$

Il faut alors estimer l'ensemble $\Phi = (\mu_1 \dots \mu_M, \Sigma_1 \dots \Sigma_M, w_1 \dots w_M)$ des paramètres du mélange. L'algorithme espérance-maximisation (*EM*) permet d'estimer l'ensemble Φ des paramètres par maximisation de la log-vraisemblance $L(P_{visage}, \Phi)$ exprimée comme il suit :

$$L(P_{visage}, \Phi) = \sum_{k=1}^N \log \left(\sum_{i=1}^M w_i G_i(\hat{u}_k) \right) \quad (24)$$

L'algorithme suppose qu'on dispose d'une pré-classification Z des données d'entrée telle que, z_{ki} vaut 1 si le stimulus \hat{u}_k appartient à la classe modélisée par G_i . Etant donné l'ensemble des paramètres courants à l'étape s de l'algorithme $\Phi^{(s)} = (\mu^{(s)}_1 \dots \mu^{(s)}_M, \Sigma^{(s)}_1 \dots$

$\Sigma^{(s)}_M, w^{(s)}_1 \dots w^{(s)}_M$, on peut estimer la probabilité conditionnelle que $z_{ki}=1$ sachant $\hat{u}=\hat{u}_k$ pour la classe i de la manière suivante :

$$t_{ki}^{(s)} = P(z_{ki} = 1 | \hat{u} = \hat{u}_k; \Phi^{(s)}) = \frac{w_i^{(s)} G_i^{(s)}(\hat{u}_k)}{\sum_{l=1}^M w_l^{(s)} G_l^{(s)}(\hat{u}_k)} \quad (25)$$

On peut alors déduire les nouveaux paramètres $\Phi^{(s+1)} = (\mu^{(s+1)}_1 \dots \mu^{(s+1)}_M, \Sigma^{(s+1)}_1 \dots \Sigma^{(s+1)}_M, w^{(s+1)}_1 \dots w^{(s+1)}_M)$ de la manière suivante :

$$\left\{ \begin{array}{l} w_i^{(s+1)} = \frac{1}{N} \sum_{k=1}^N t_{ki}^{(s)} \\ \mu_i^{(s+1)} = \frac{\sum_{k=1}^N t_{ki}^{(s)} \hat{u}_k}{\sum_{k=1}^N t_{ki}^{(s)}} \\ \Sigma_i^{(s+1)} = \frac{\sum_{k=1}^N t_{ki}^{(s)} (\hat{u}_k - \mu_i^{(s+1)})^T (\hat{u}_k - \mu_i^{(s+1)})}{\sum_{k=1}^N t_{ki}^{(s)}} \end{array} \right. \quad (26)$$

Le processus est itéré jusqu'à ce que l'amélioration de la vraisemblance logarithmique soit inférieure à un seuil fixé manuellement, ou tant que le nombre d'itérations est inférieur à un seuil fixé à l'avance. Une fois les paramètres du mélange de gaussiennes estimés, on peut alors calculer les probabilités d'appartenance des pixels aux différentes classes. Les pixels sont associés à la classe pour laquelle la probabilité d'appartenance est la plus grande :

$$P(\hat{u}_k \in Y_i) = \frac{w_i G_i(\hat{u}_k)}{\sum_{l=1}^M w_l G_l(\hat{u}_k)} \quad (27)$$

2.5.2.2 Initialisation de l'algorithme Espérance-Maximisation : Algorithme de K-moyennes

Pour initialiser l'algorithme *EM* nous avons vu qu'il faut disposer d'une pré-classification des données d'entrée ainsi que des estimations des moyennes et des variances des classes. L'algorithme *EM* étant un algorithme d'estimation itératif, plus les paramètres initiaux seront proches du résultat final, plus la convergence sera rapide. Nous avons utilisé

l'algorithme des K-moyennes pour effectuer la pré-classification des données d'entrée. L'algorithme des K-moyennes est doublement intéressant dans notre cas. Il permet de trouver une partition en M classes $\{Y_1 \dots Y_M\}$ des données d'entrée et de trouver les M centres $C = \{c_1 \dots c_M\}$ des classes. Cette partition Q est dite rigide, c'est-à-dire que chaque stimulus d'entrée est associé à une seule classe :

$$Q = \begin{bmatrix} q_{1,1} & \cdots & q_{1,N} \\ \vdots & \ddots & \vdots \\ q_{M,1} & \cdots & q_{M,N} \end{bmatrix} \quad (28)$$

Avec $q_{ij} = 1$ quand le stimulus \hat{u}_j appartient à la classe Y_i . De plus, on a les contraintes suivantes sur la partition :

$$\sum_{i=1}^M q_{i,j} = 1, \quad j = 1, \dots, N \quad (29)$$

$$\sum_{j=1}^N q_{i,j} > 1, \quad i = 1, \dots, M \quad (30)$$

Connaissant les centres $\{\mu_1 \dots \mu_M\}$, la fonction objective à minimiser est de la forme suivante :

$$F(U, C) = \sum_{j=1}^N \sum_{i=1}^M (q_{i,j}) \|\hat{u}_j - c_i\|^2 \quad (31)$$

Les étapes de l'algorithme des K-moyennes sont alors :

- Initialisation des centres $C(0)$ en choisissant M stimuli parmi les N données d'entrée.
- Déterminer la partition initiale Q :

$$q_{i,j} = \begin{cases} 1 & \text{si } \|\hat{u}_j - c_i\| = \min_k \|\hat{u}_j - c_k\|, \\ 0 & \text{autrement} \end{cases}, \quad (32)$$

$i = 1, \dots, M, j = 1, \dots, N$

- On calcule ensuite les nouveaux centres $C(t)$:

$$c_i = \frac{\sum_{j=1}^N q_{i,j} \hat{u}_j}{\sum_{j=1}^N q_{i,j}} \quad (33)$$

- On calcule la nouvelle partition $Q(t)$ à l'aide de (32).
- On répète tant que $Q(t) \neq Q(t-1)$ et $t < t_{max}$.

Les centres $C = \{c_1, \dots, c_M\}$ ainsi trouvés, sont utilisés comme les moyennes initiales des gaussiennes G modélisant les M classes. Les variances $\{\Sigma_1, \dots, \Sigma_M\}$ sont calculées à partir des centres C et de la partition Q . Enfin, la pré-classification Z est initialisée avec Q .

2.5.2.3 Identification des pixels de visage sur des images de bouche

Nous avons fait l'hypothèse que les images traitées dans ce chapitre étaient cadrées sur la moitié basse du visage. Ceci implique que la plus grande partie des pixels de l'image appartient au visage. Soit l'ensemble $P = [\hat{u}_1, \dots, \hat{u}_N]$ des pixels de l'image de bouche. P est constitué des valeurs des pixels dans \hat{U} . Pour une valeur de M , la distribution G_i pour laquelle w_i est maximum, ainsi que les distributions G_l , dont la moyenne μ_l est inférieure à μ_i , seront associées aux pixels du visage d'après notre hypothèse sur le cadrage. Le problème à résoudre est celui du nombre de classes M considéré pour l'estimation du mélange de gaussiennes et l'initialisation de l'algorithme EM .

Suivant la qualité du cadrage de l'image, il pourra se trouver des pixels n'appartenant pas au visage dans l'image traitée (pixels de l'arrière plan, des vêtements, ...). Nous avons donc supposé que les pixels d'une image de bouche pouvaient toujours être séparés en au moins 2 classes de pixels : une classe de pixels du visage et une classe de pixels n'appartenant pas au visage. Selon les images, l'état de la bouche pourra varier et on peut voir apparaître l'intérieur de la bouche, des dents, la langue ou des combinaisons de ces différents éléments. Il est donc nécessaire de posséder un critère permettant de déterminer

le nombre optimal de classes pour séparer les pixels de l'image d'entrée. L'algorithme *EM* ne permet pas de garantir que les variances interclasses des gaussiennes du mélange soient maximales et que les variances intraclasses soient minimales. Dans notre approche, nous considérons que les gaussiennes qui composent le mélange modélisent les classes des pixels de l'image. On va donc chercher à obtenir la meilleure séparation entre les classes de pixels. Cela revient à déterminer un mélange de gaussiennes avec des variances interclasses grandes et des variances intraclasses faibles.

Pour une image d'entrée inconnue, on initialisera le nombre de gaussiennes du mélange à $M = 2$. Comme nous supposons que la majorité des pixels de l'image appartient au visage, on effectuera l'estimation des mélanges de gaussiennes pour M croissant, avec un pas unitaire, tant que le paramètre w_{max} maximal du mélange avec M gaussiennes est supérieur à 0.5. Pour chaque valeur de M , l'algorithme *EM* s'arrête lorsque l'amélioration de la log-vraisemblance est inférieure à 1%. Le mélange pour lequel le rapport V_{intra}/V_{inter} est minimal, et avec $w_{max} > 0.5$, sera considéré pour la classification des pixels du visage. De cette manière, on restreindra le nombre de classes possible. Avec un trop grand nombre de classes, on risquerait une sur-segmentation du visage. La table 2.3 présente les variances intraclasses, interclasses, le rapport V_{intra}/V_{inter} ainsi que le paramètre w_{max} de la classe prépondérante du mélange avec $M=[2,3,4,5]$ pour l'image de bouche de la figure 2.10-a. Pour cette image l'algorithme s'arrête pour $M = 5$ et le nombre optimal de classes est $M = 4$ d'après nos critères. On donne également sur la figure 2.10 les histogrammes de teinte et les tracés des gaussiennes estimées qui composent les mélanges pour $M=1,\dots,4$.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}	w_{max}
$M=2$	347.1	748.1	0.46	0.76
$M=3$	633.7	461.5	1.37	0.58
$M=4$	165.9	929.3	0.17	0.63
$M=5$	316.4	2928.4	0.11	0.42

Table 2.3 : Variance intraclasses, Variance interclasses et V_{intra}/V_{inter} ainsi que le paramètre w_{max} maximum des mélanges avec $M=[2,3,4,5]$ gaussiennes.

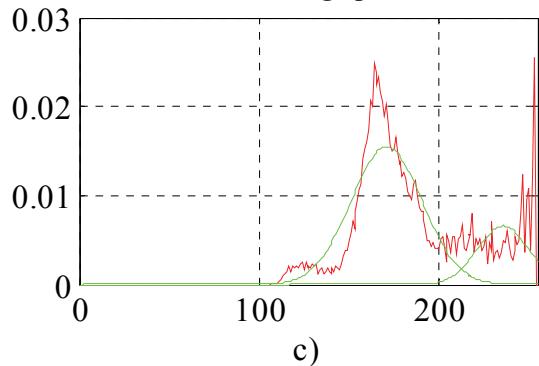


a)



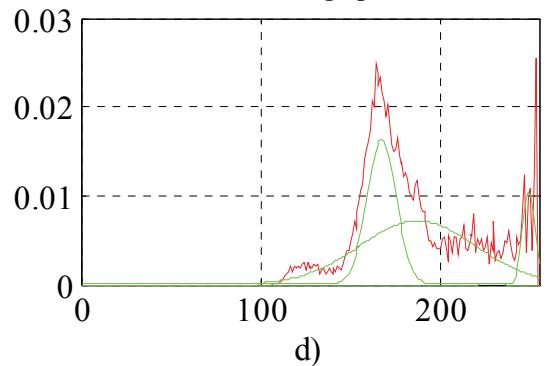
b)

Tracé des distributions gaussiennes issues du mélange pour $M=2$



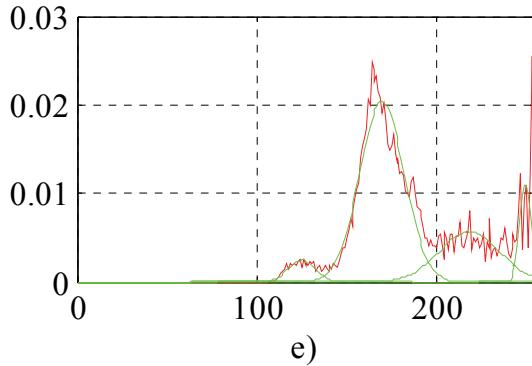
c)

Tracé des distributions gaussiennes issues du mélange pour $M=3$



d)

Tracé des distributions gaussiennes issues du mélange pour $M=4$



e)

Figure 2.10 : Tracés des mélanges de gaussiennes pour $M=[2,3,4]$: a) Image d'entrée, b) Image de teinte \hat{U} , c) Mélange pour $M=2$, d) Mélange $M=3$, e) Mélange pour $M=4$, en rouge, on donne l'histogramme de l'image dans \hat{U} , en vert, on donne les tracés des distributions gaussiennes issues des mélanges.

D'après la figure 2.10, $M=2$ gaussiennes semble insuffisant pour modéliser la distribution de teinte de l'image. Les zones du visage et de l'arrière plan sont grossièrement modélisées. On voit, sur la figure 2.11-a, qu'on détecte, néanmoins, bien le visage mais que l'on détecte également des pixels appartenant aux vêtements. Pour $M=3$, on constate que les

gaussiennes issues du mélange se recouvrent. Ceci est dû à l'algorithme *EM* qui estime les paramètres du mélange au sens du maximum de vraisemblance. Il n'existe aucune contrainte sur les variances intraclasses et interclasses. Les distributions gaussiennes peuvent se recouvrir comme dans notre exemple avec $M=3$. Avec notre exemple, pour $M=3$, en effectuant une classification basée sur la probabilité d'appartenance pour la distribution gaussienne correspondant au poids w_i le plus grand, les pixels du visage sont exclus (figure 2.11-b).

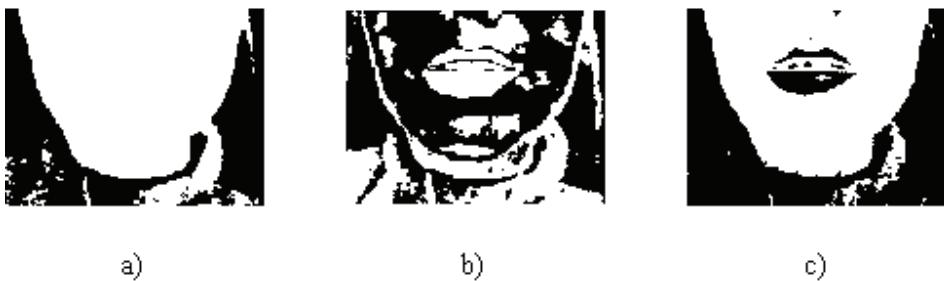


Figure 2.11 : Résultat de la détection de la zone du visage pour $M=[2,3,4]$

La meilleure modélisation dans ce cas est obtenue pour $M=4$. Sur le tracé des distributions gaussiennes issues de l'estimation du mélange, on remarque que les distributions gaussiennes modélisent bien l'ensemble de l'histogramme de \hat{U} , le pic le plus grand correspondant aux pixels de peau. Lorsque l'on classe les pixels en calculant leurs probabilités d'appartenances, on obtient le masque de la figure 2.11-c.



Figure 2.12 : Exemple de détection des pixels du visage.

Sur la figure 2.11-c, la zone du visage est bien identifiée et le nombre d'erreurs est réduit. On remarque cependant que, dans ce cas, les pixels des lèvres sont exclus du masque, car le

contraste est fort avec les pixels de peau. Nous avons vu que, dans \hat{U} , la moyenne des pixels des lèvres était inférieure à la moyenne des pixels de peau. Il faut ajouter au masque binaire du visage les pixels correspondant aux classes dont la moyenne est inférieure à celle de la classe visage, si ces classes existent. Pour le cas de l'image de la figure 2.10-a, on obtient le masque de la figure 2.12 qui englobe l'ensemble des pixels du visage.

2.5.3 Localisation et segmentation des lèvres

2.5.3.1 Segmentation région-contour des lèvres

Dans la section précédente, nous nous sommes intéressés à l'identification des pixels appartenant à la zone du visage. Le but de ce prétraitement est d'éliminer les pixels parasites afin d'exploiter au mieux la séparation entre les pixels des lèvres et de la peau dans \hat{U} . On suppose, à partir de maintenant, que l'on dispose d'un masque du visage, comme dans le cas de la figure 2.12. On va maintenant chercher à isoler les lèvres du reste du visage.

À partir des pixels inclus dans le masque du visage obtenu précédemment, nous allons effectuer une recherche de la zone de la bouche par une méthode combinée région-contour. Nous supposons que la bouche est incluse strictement dans la zone du visage. Dans la teinte \hat{U} , la moyenne de la classe des pixels des lèvres est inférieure à la moyenne de la classe des pixels de la peau. A priori, les pixels les plus sombres du masque du visage correspondront aux pixels des lèvres. Nous proposons d'appliquer un algorithme de seuillage automatique basé sur cette propriété pour localiser les lèvres rapidement.

Les étapes de l'algorithme sont les suivantes :

- On pose $\varepsilon = 0$.
- Soit μ_{visage} la moyenne dans \hat{U} de l'ensemble des pixels appartenant au masque du visage.
- Soit \hat{u}_{min} le niveau de teinte minimale dans \hat{U} sur l'ensemble des pixels du masque du visage.
- Soit $se = \mu_{visage}$ le seuil initial, alors :
 - On détermine le masque binaire composé des pixels de visage dont le niveau de teinte \hat{u} est tel que : $\hat{u} < se$ et $\hat{u} > \hat{u}_{min}$.

-
- On effectue, ensuite, une labellisation du masque binaire et on ne garde que les blobs qui sont strictement inclus dans le masque du visage.
 - On détermine le plus petit polygone convexe P_{CONV} englobant les blobs résultants et on calcule la somme des flux $F(se)$ des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ pour $t=[1, \dots, N_{echelle}]$ au travers de P_{CONV} .
 - On calcule la variance interclasses et la variance intraclasses des classes de pixels des lèvres et de la peau ainsi que le rapport V_{intra}/V_{inter} .
 - On détermine le critère de performance $CP(se)=F(se)/(V_{intra}/V_{inter})$.
 - On pose $se = se - 1$ et on répète les 5 points précédents tant que $se > \hat{u}_{min}$.
 - On cherche, ensuite, le seuil se pour lequel CP est maximal et le masque correspondant est considéré comme le masque candidat des lèvres.
 - On pose $\varepsilon = \varepsilon + 1$, on réduit la zone de recherche aux limites haute et basse du masque candidat des lèvres et on répète l'algorithme de seuillage automatique jusqu'à ce que l'erreur entre 2 masques candidats successifs, à $\varepsilon-1$ et à ε , soit inférieure à 1 % de la taille du masque obtenu en $\varepsilon-1$.

Pour localiser la bouche sur la zone du visage, nous avons privilégié une approche par seuillage automatique. Nous savons que les niveaux de teinte des pixels des lèvres dans \hat{U} sont inférieurs à ceux des pixels de peau. Comme nous avons identifié la zone du visage, a priori, les pixels les plus sombres dans le masque du visage correspondent aux pixels des lèvres. Dans le cas de la recherche du visage, nous avons appliqué une méthode basée sur l'estimation d'un mélange de gaussiennes en recherchant le nombre optimal de classes à l'aide d'un critère statistique (le rapport entre variance intraclasses et variance interclasses) et d'un critère morphologique (la taille minimale de l'ensemble des pixels de visage). Dans le cas de la recherche des lèvres, il est difficile de faire une hypothèse sur la taille minimale de l'ensemble des pixels des lèvres. L'épaisseur des lèvres peut varier fortement d'un individu à un autre, tout comme l'état de la bouche. Nous avons donc privilégié l'approche par seuillage, dans laquelle nous n'avons pas d'a priori sur la surface de l'ensemble des pixels des lèvres. Nous nous basons uniquement sur les niveaux de teinte des pixels des lèvres et sur l'information des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$.

A l’itération initiale $\varepsilon = 0$ de l’algorithme de seuillage, la zone de recherche correspond au masque du visage obtenu à l’étape précédente (figure 2.13-c). Le seuil de départ se correspond à la moyenne μ_{visage} dans \hat{U} de l’ensemble des pixels appartenant au masque du visage. On effectue un seuillage sur les pixels du masque de visage en ne gardant que les pixels inférieurs à se . On ajoute alors la contrainte géométrique que le masque de la bouche est strictement inclus dans le masque du visage. On effectue alors une labellisation du masque binaire candidat, et on ne garde que les blobs qui sont strictement inclus dans le masque du visage. Les figures 2.13-d et 2.13-e illustrent des cas de figure possibles. Sur la figure 2.13-d, le seuil est trop haut, le blob contenant la zone des lèvres n’est pas strictement inclus dans la zone du visage, on ne garde alors que le blob en blanc qui correspond à une zone sur le menton. Sur la figure 2.13-e, le seuillage est effectué avec une valeur du seuil pertinente. Le masque binaire candidat est composé d’un blob correspondant aux lèvres et d’un blob situé sur la limite haute de la zone du visage. Cette fois ci, on ne gardera que le blob correspondant aux lèvres.

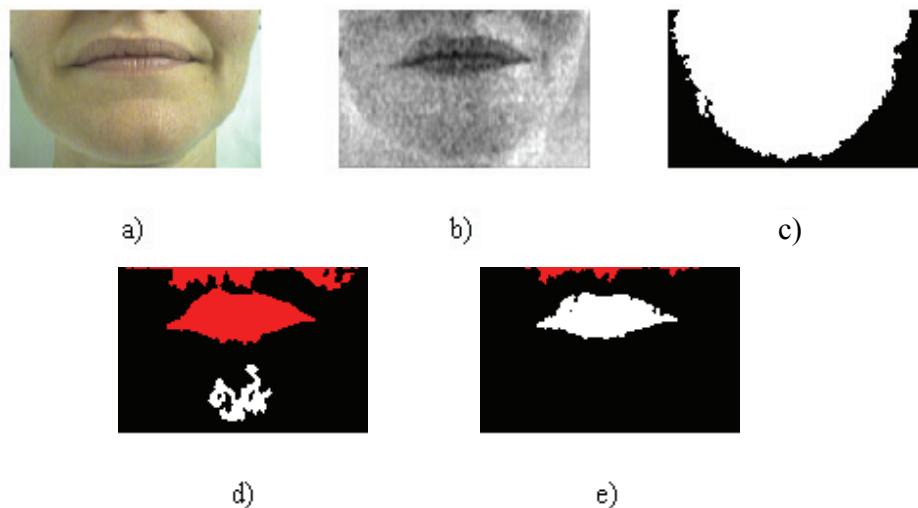


Figure 2.13 : Exemples de masques candidats des lèvres pour une image de bouche, a) Image d’entrée, b) teinte \hat{U} , c) masque du visage, d) masque candidat des lèvres pour un seuil trop élevé, e) masque candidat des lèvres pour une valeur de seuil pertinente, les zones en rouge ne sont pas strictement incluses dans le masque du visage, les zones en blanc sont strictement incluses dans le visage.

Une fois que l’on a appliqué la contrainte géométrique sur le masque binaire, on détermine le plus petit polygone convexe $PCONV$ englobant les blobs restant et on calcule la somme

des flux des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ pour $t=[1, \dots, N_{echelle}]$. La figure 2.14 présente le tracé du polygone convexe $PCONV$ entourant le masque binaire candidat des lèvres pour le cas du masque de la figure 2.13-e.



Figure 2.14 : Tracé du plus petit polygone convexe englobant le masque candidat des lèvres de la figure 2.13-e.

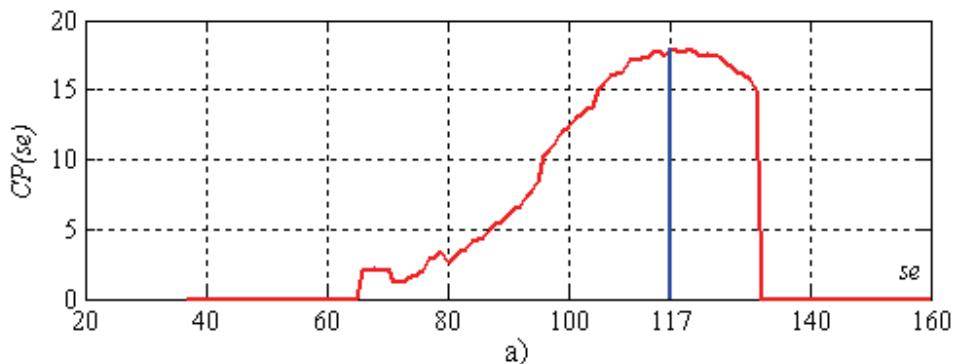
La somme $F(se)$ des flux du gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ au travers du polygone $PCONV$ à l'échelle t , pour $t=[1, \dots, N_{echelle}]$ s'exprime de la manière suivante :

$$F(se) = \sum_{t=1}^{N_{echelle}} \int_{PCONV} \nabla_{norm}L(x, y, t) dn \quad (34)$$

où dn est orthogonal au contour $PCONV$. $N_{echelle}$ correspond à l'échelle la plus petite (avec le moins de détails) que l'on considère dans notre famille de gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ (voir section 2.4). Le choix de $N_{echelle}$, et donc du nombre d'échelles considérées dans le calcul de $F(se)$, fera l'objet de la section suivante. Avec cette expression du flux, plus l'intensité du gradient sera importante le long du contour $PCONV$, plus la somme $F(se)$ sera grande, et plus le périmètre du contour sera grand, plus $F(se)$ sera grande. Le but est de défavoriser les blobs parasites de petite taille sur lesquels le calcul d'un flux peut être très grand. On calcule ensuite la variance interclasses et la variance intraclasses des classes des pixels des lèvres et de la peau. Pour finir on calcule le critère de performance $CP(se)=F(se)/(V_{intra}/V_{inter})$. Par la suite, on pose $se = se - 1$ et on répète les étapes de seuillages tant que $se > \hat{u}_{min}$. On donne, sur la figure 2.15-a, le tracé de $CP(se)$ pour l'image de bouche de la figure 2.13-a. Pour cette image de test, $\mu_{visage} = 161$ et $\hat{u}_{min} = 37$ dans \hat{U} et $N_{echelle}=(3)^2$. Le seuil maximisant $CP(se)$ est $se_{opt}=117$. La figure 2.15-b présente le masque candidat des lèvres correspondant à ce seuillage. La conjugaison des

informations de teinte et des gradients permet de sélectionner un masque dont les contours correspondent à des gradients forts.

Après détermination du seuil optimal avec comme zone de recherche le masque du visage, on obtient alors un masque des lèvres (figure 2.15-b). On passe à l'itération $\varepsilon=\varepsilon+1$. On extrait les limites haute et basse du masque des lèvres obtenu à la phase précédente et on recadre la zone de recherche des lèvres entre ces limites (figure 2.16).



b)

Figure 2.15 : Tracé de $CP(se)$ avec $N_{echelle} = (3)^2$ pour l'image de bouche de la figure 2.13-a, a) tracé de $CP(se)$, b) masque candidat des lèvres pour le seuil maximisant $CP(se)$.

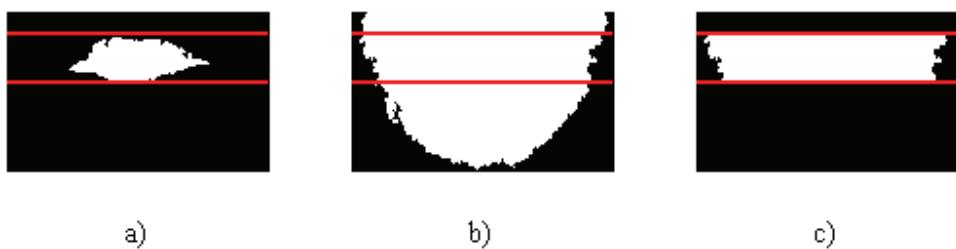


Figure 2.16 : Recadrage de la zone de recherche de la bouche sur la zone du visage, a) masque candidat des lèvres ainsi que le tracé des limites haute et basse, b) masque du visage, c) nouvelle zone de recherche.

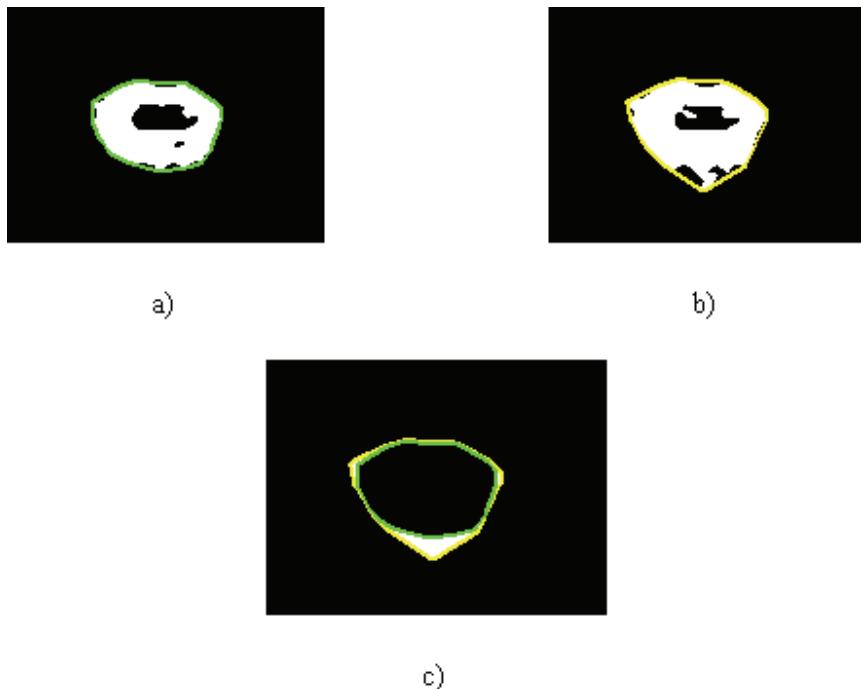


Figure 2.17 : Aire entre le contour convexe de deux masques des lèvres candidats, a) masque candidat à $\varepsilon-1$, b) masque candidat à ε , c) aire entre les 2 masques.

On répète les opérations de recherche du masque des lèvres dans la nouvelle zone de recherche. Si l'aire entre le contour convexe du nouveau masque et le contour convexe du masque candidat de l'itération précédente (voire figure 2.17) est inférieure à 1% de l'aire du masque candidat précédent, on arrête la recherche des lèvres et le dernier masque candidat est considéré comme le masque des lèvres. Autrement, on continue la recherche.

2.5.3.2 Choix de $N_{echelle}$

Les contours de la bouche sont caractérisés par des maximums locaux de l'intensité des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$. Dans la section 2.4 de ce chapitre, nous avons étudié l'intérêt d'utiliser une représentation multi-échelle des gradients pour caractériser les contours de la bouche et en particulier le contour externe. Dans cette étude, nous avons mis en évidence, d'une part, que l'échelle de l'image d'entrée n'était pas nécessairement optimale pour caractériser le contour externe de la bouche et, d'autre part, que les échelles optimales n'étaient pas forcément les mêmes pour le contour externe supérieur et pour le contour externe inférieur. D'une manière générale, on peut tout de même énoncer

l'hypothèse suivante : il existe des échelles pour lesquelles l'intensité des gradients γ -normalisés $\nabla_{norm}L(x,y,t)$ admet un maximum local par rapport aux échelles pour le contour externe de la bouche. Il faut aussi souligner le fait que dans le cas d'un sujet avec des lèvres fines, plus l'échelle est diminuée, plus il sera difficile de discriminer le contour externe et le contour interne des lèvres. Il y a donc un compromis à faire entre l'intensité des gradients le long des contours de la bouche et le niveau de détail de la forme. Idéalement, pour composer la famille de gradients γ -normalisés $\nabla_{norm}L(x,y,t)$, il faudrait partir de l'échelle d'entrée de l'image, c'est-à-dire $L(x,y,0)$ qui possède le maximum de détail, et ajouter les échelles $L(x,y,t)$ à notre famille tant que l'intensité du gradient croît sur les contours de la bouche. En pratique, nous ne connaissons pas, a priori, la position des contours de la bouche. Nous avons envisagé une approche globale tirant parti de notre connaissance de la localisation du visage. Dans notre cas, nous nous sommes intéressés à l'intensité des gradients γ -normalisés le long de la ligne verticale passant au milieu de la zone du visage (figure 2.18). On sait que la bouche se trouve au centre de la zone du visage. On fait l'hypothèse que cette ligne coupe les contours externe supérieur et externe inférieur ainsi que les contours internes de la bouche. Les orientations horizontales étant prépondérantes sur la bouche, les contours horizontaux seront, a priori, les plus forts.

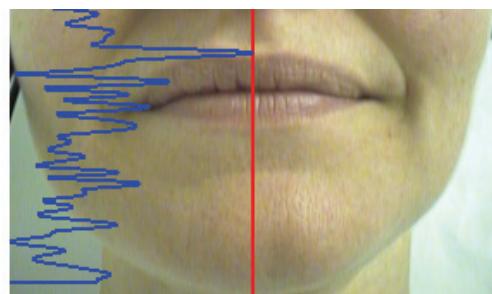


Figure 2.18 : Représentation de l'intensité du gradient γ -normalisé $\nabla_{norm}L(x,y,t)$, pour $t=1$ sur la ligne verticale passant au milieu de la zone du visage

On constate, sur la figure 2.18, que les contours de la bouche, en particulier les contours externes supérieur et inférieur, sont bien caractérisés par des maximums locaux de l'intensité. Etant donné que l'on ne connaît pas les positions spatiales des maxima locaux correspondant aux contours de la bouche, notre approche pour construire la famille de

gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ est de partir de l'échelle $t=\sigma^2=1$. Ensuite, on calcule la somme S des énergies de tous les maximums locaux se trouvant sur la ligne verticale et qui sont inclus dans le masque du visage. Par la suite, on pose $\sigma=\sigma+1$ et on passe à la nouvelle échelle $t=\sigma^2$. On calcule de nouveau la somme des énergies des gradients γ -normalisés aux mêmes positions qu'à l'échelle $t=1$. Si cette somme est supérieure, on ajoute cette échelle à notre famille et on continue en incrémentant σ de 1, sinon l'ajout d'échelle est arrêté. De cette manière, on déterminera l'échelle pour laquelle globalement l'intensité des gradients γ -normalisés est maximale pour la zone de la bouche. Pour le cas de l'image de bouche de la figure 2.13-a, nous avons tracé l'évolution de la somme des énergies des maximums locaux en fonction de l'échelle $t=\sigma^2$ ainsi que les images d'énergie des gradients γ -normalisés aux échelles correspondantes (figure 2.19).

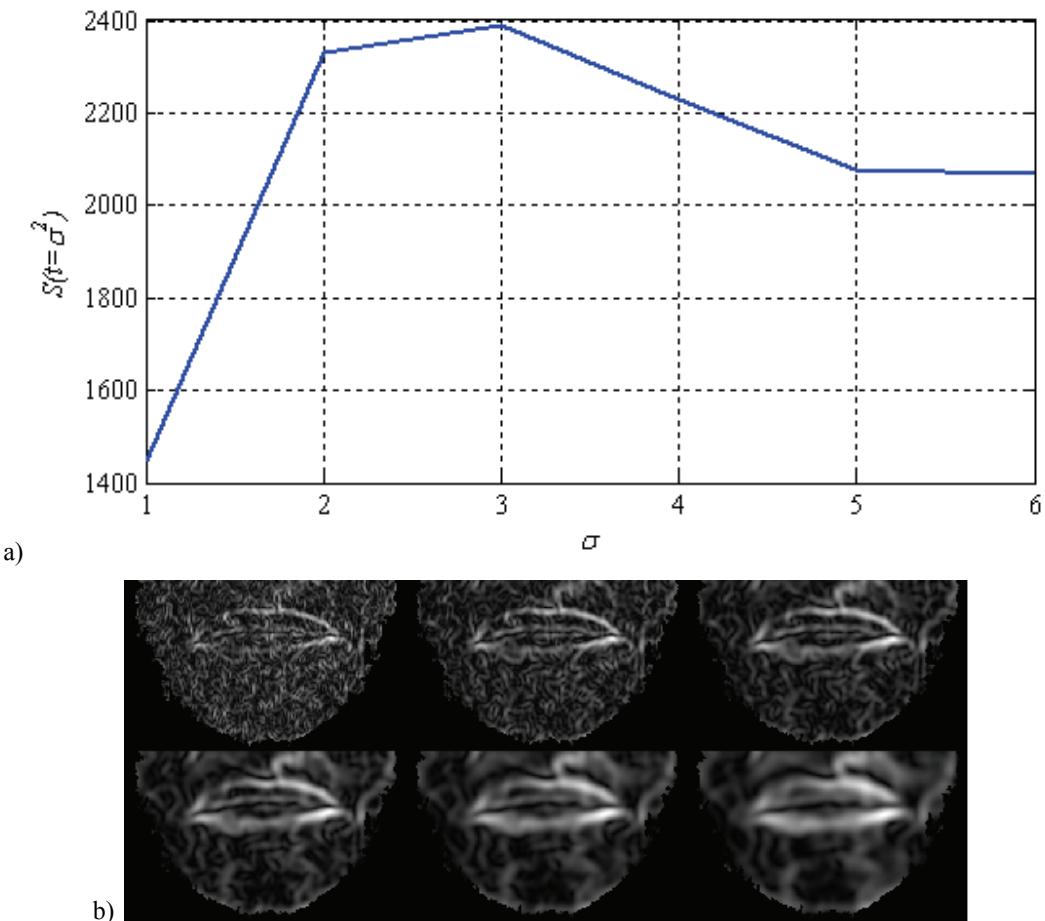


Figure 2.19 : Évolution de $S(t=\sigma^2)$, a) tracé de $S(t=\sigma^2)$, b) images de l'intensité des gradients $\nabla_{norm}L(x, y, t)$ pour $1 < \sigma < 6$.

Pour ce cas, on constate qu'à partir de $\sigma=3$, $S(t=\sigma^2)$ décroît. On obtient alors que $N_{echelle} = (3)^2$. On ne gardera que les $\nabla_{norm}L(x, y, t)$ pour $t < (3)^2$. On voit que pour $\sigma=3$ la saillance du contour externe de la bouche est bien meilleure que pour $\sigma=1$. On constate également que le contour externe est encore bien détaillé. Il est également intéressant de rappeler que les gradients $\nabla_{norm}L(x, y, t)$ sont issus de la représentation multi-échelle $L(x, y, t)$ de l'image d'entrée (voir section 2.4). En pratique cette représentation correspond au filtrage de l'image par des noyaux gaussiens d'écart type σ croissant. On peut interpréter $N_{echelle}$ comme une estimation de la largeur, en pixels, des contours prépondérants de la zone du visage.

2.6 Bilan

Les figures 2.20 et 2.21 présentent des exemples de segmentation des lèvres. À la figure 2.22, nous donnons également des exemples de segmentation des lèvres pour des sujets ayant la peau noire extraits de la base FERET (Philipps, 2000). On constate que la méthode de segmentation fonctionne également pour les sujets de couleur de peau noire.



Figure 2.20 : Exemples de segmentations des lèvres.



Figure 2.21 : Exemples de segmentations des lèvres.



Figure 2.22 : Exemples de segmentations des lèvres pour des sujets ayant la peau noire.

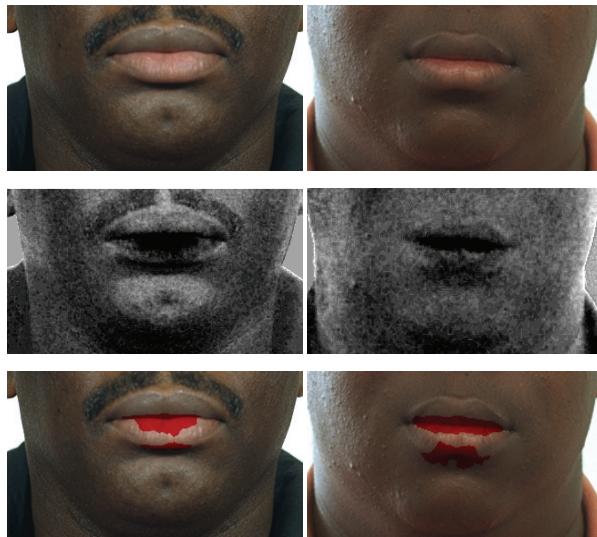


Figure 2.23 : Exemples de segmentations erronées.

La principale difficulté que nous avons rencontrée pour segmenter les lèvres dans le cas de sujet ayant la peau noire vient de l'absence de contraste entre les lèvres et la peau et de l'inhomogénéité de teinte des pixels des lèvres. A la figure 2.23, nous présentons des cas pour lesquels la segmentation des lèvres a échoué. On peut voir sur la deuxième ligne les images de teinte \hat{U} . On constate que, malgré le traitement par l'algorithme allongement-décorrélation, il n'y pas de contraste entre la peau et la lèvre supérieure et que les niveaux de teinte des 2 lèvres sont différents. Dans nos hypothèses, nous avons supposé que les lèvres constituaient un ensemble de pixels homogène et séparable de l'ensemble des pixels de la peau. La segmentation de la lèvre supérieure dans les cas de la figure 2.23 sera donc très difficile. Il est à noter que l'on parvient, néanmoins, à localiser grossièrement la zone de la bouche. Le faible nombre d'images de bouche de sujet ayant la peau noire dont nous disposions ne nous a pas permis de faire une étude approfondie. Il reste que, dans ces cas de figure, la teinte semble insuffisante pour discriminer les lèvres. La constitution d'une base d'images composée de sujets de couleur de peau noire sera un objectif pour la suite de nos travaux. Du point du vue de la complexité, le temps de calcul pour déterminer le masque binaire des lèvres avec une image de bouche d'une résolution de 102x170 pixels est de 5s. La machine utilisée était équipée d'un processeur double cœur avec une fréquence d'horloge de 2.33GHz. Nos algorithmes ont été codés sous Matlab. En ce qui concerne l'évaluation de la segmentation des lèvres, nous effectuerons une évaluation sur la

modélisation des contours au chapitre 5. L'évaluation de la segmentation des lèvres sera incluse dans l'évaluation la segmentation des contours des lèvres.

Chapitre 3. Détection de l'état de la bouche

3.1 Introduction

Dans le chapitre précédent, nous avons décrit notre méthode permettant de localiser la bouche sur des images de visage et d'extraire un masque binaire des lèvres. A partir de maintenant, nous supposons que nous avons localisé la région de la bouche à partir des traitements du chapitre 2. Cette segmentation, qui nous permet d'identifier grossièrement le contour externe de la bouche, n'est pas suffisamment robuste, ni précise, pour détecter le contour interne. En effet, la configuration de la région interne de la bouche peut être très variable et perturber la modélisation des contours internes de la bouche. Lorsque celle-ci est fermée, le contour interne se résume à la jointure entre les 2 lèvres. Lorsque la bouche est ouverte, il y aura, comme pour le contour externe, un contour interne supérieur et un contour interne inférieur. Pour modéliser de manière robuste les contours de la bouche, externes ou internes, l'initialisation est déterminante. La connaissance de l'état de la bouche (ouvert ou fermé) sera une information importante pour modéliser les contours internes de la bouche. A la section 1.5.2.3, nous avons présenté plusieurs méthodes proposées pour déterminer l'état de la bouche dans l'optique d'initialiser un modèle paramétrique. La plupart du temps, les auteurs utilisent les informations de gradient ou de couleur pour déterminer l'état de la bouche. Récemment dans (Benoit, 2005), l'auteur a utilisé une méthode fréquentielle basée sur un modèle de Système Visuel Humain (*SVH*) pour déterminer l'état dynamique de la bouche sur des séquences vidéo. Le but était de prévenir l'état d'hypovigilance d'une personne au volant d'une voiture. Dans le cadre de nos travaux, nous souhaitions pouvoir identifier l'état de la bouche sur des images de visage statiques. Nous développerons dans ce chapitre les étapes de la méthode d'identification de l'état de la bouche que nous avons développée en nous inspirant des modèles du *SVH* existants.

La section 3.2 sera consacrée à l'introduction des modèles inspirés par le *SVH* pour traiter des images statiques. Un modèle de rétine, permettant de renforcer les contours des images, sera d'abord décrit. Ensuite, nous nous intéresserons au modèle de cortex visuel. Ce modèle conduit au calcul d'un spectre Log-polaire contenant l'information de forme et d'apparence de l'image. Dans la section 3.3, nous présenterons la méthode d'identification de l'état de la

bouche basée sur un classifieur supervisé. Enfin, la section 3.4 présentera les résultats expérimentaux.

3.2 Modélisation du système visuel humain pour le traitement d'image statique

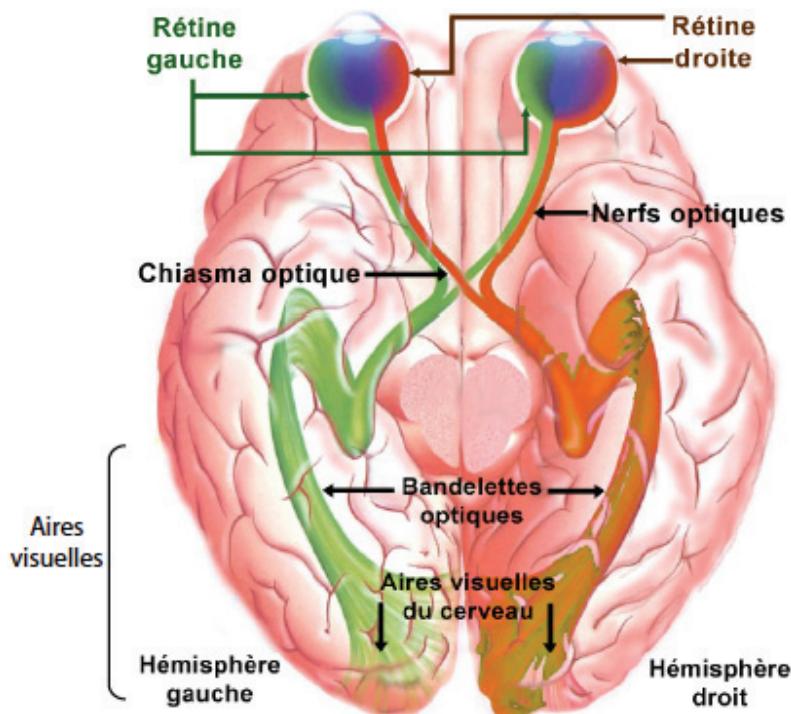


Figure 3.1 : Schéma du système visuel humain (Benoit, 2007).

En vision, lorsque l'on s'attaque à un problème particulier, la démarche classique consiste à appliquer des outils mathématiques pour modéliser un phénomène, une action ou pour reproduire un traitement particulier qui est ensuite implémenté sous forme d'algorithme. Pour ce qui est de la vision, une autre démarche consiste à s'inspirer du vivant. Pour un humain il est tout à fait naturel d'analyser et d'interpréter une image quelconque, ce qui est, à l'heure actuelle, encore impossible, même pour les algorithmes de vision les plus puissants. Le système visuel humain est de loin le système de vision le plus performant connu. La figure 3.1 présente un schéma global du système visuel humain. L'information

visuelle est reçue par l'œil puis est ensuite transmise au cerveau. La rétine effectue des traitements préliminaires pour convertir l'excitation lumineuse en signaux qui seront, ensuite, transmis au cerveau par le nerf optique. Les traitements de haut niveau seront réalisés dans le cerveau, au niveau des aires visuelles qui se trouvent à l'arrière des 2 hémisphères du cerveau. Dans la suite de cette section, nous introduirons les modèles de rétine et de cortex visuel que nous avons utilisés pour le problème de l'identification de l'état de la bouche. Pour une étude complète du traitement de l'information dans la rétine chez les vertébrés, les lecteurs pourront se référer à (Hérault, 2001). On rappelle que dans le cadre de l'identification de l'état de la bouche, nous avons fait l'hypothèse que nous travaillons sur des images de bouche statiques, nous nous limiterons donc à l'étude des traitements spatiaux du système visuel humain. De plus, dans l'exposé sur les modèles de rétine et de cortex visuel, l'étude concerne l'analyse de la luminance, par souci de simplification.

3.2.1 Rétine

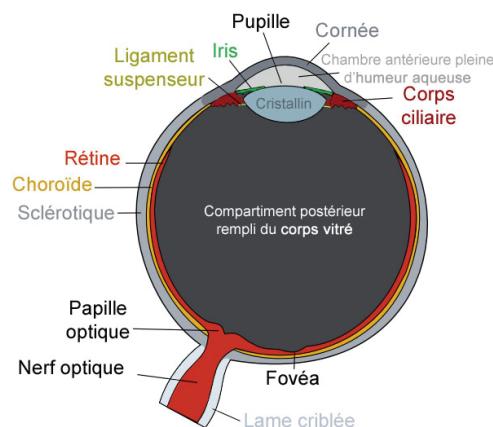


Figure 3.2 : Schéma de l'œil humain (http://fr.wikipedia.org/wiki/Oeil_humain).

Le système d'acquisition du SVH est constitué par l'œil. La figure 3.2 présente un schéma de l'œil humain. Les principaux éléments de l'œil sont :

- La cornée, membrane transparente, qui forme une barrière protectrice entre l'intérieur de l'œil et l'extérieur.

- L’iris et le cristallin qui constituent le système optique de l’œil. L’iris, dont le diamètre est variable, permet d’ajuster la quantité de lumière qui entre dans l’œil tandis que le cristallin permet d’ajuster la focale de l’œil.
- La rétine tapisse le fond de l’œil. Elle constitue le capteur du *SVH*.

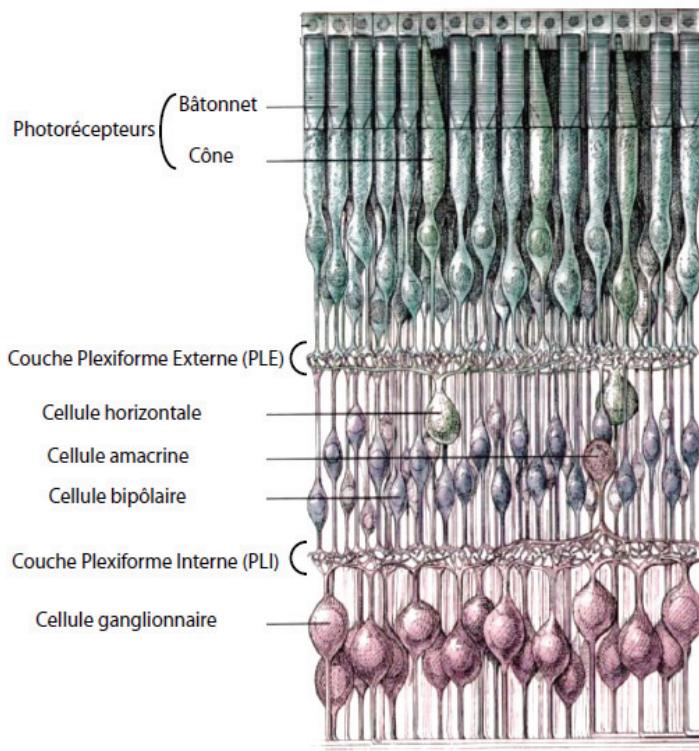


Figure 3.3 : Organisation de la rétine (Benoit, 2007)

La rétine est l’élément qui convertit l’excitation lumineuse reçue par l’œil en signaux électriques qui sont ensuite convoyés vers le cerveau par le nerf optique. En ce sens, elle est comparable au capteur d’un système d’acquisition tel que le capteur CCD ou CMOS d’une caméra. La rétine est formée d’un assemblage de cellules spécialisées organisées en 3 couches (figure 3.3):

- La couche des photorécepteurs.
- La couche des cellules horizontales, bipolaires et amacrines.
- La couche des cellules ganglionnaires.

La zone d'interconnexion, entre la couche des photorécepteurs et la couche des cellules horizontales, bipolaires et amacrines, s'appelle la couche PLexiforme Externe (*PLE*). La zone d'interconnexion, entre les cellules horizontales et bipolaires et la couche des cellules ganglionnaires, s'appelle la couche PLexiforme Interne (*PLI*).

3.2.1.1 Photorécepteurs

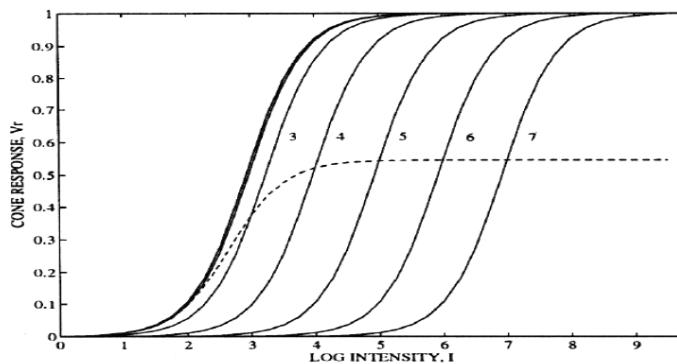


Figure 3.4 : Réponse d'un photorécepteur en fonction de la lumière reçue dans son voisinage (Kolb, 1996). En trait plein sont tracées les dynamiques d'un photorécepteur en fonction de la luminance locale. En pointillé est tracée la moyenne de la dynamique.

Les photorécepteurs constituent la première couche de la rétine. Leur rôle est de coder l'excitation lumineuse sous forme d'un potentiel de membrane. Les photorécepteurs, dont la dynamique ne dépasse pas 1 à 2 décades, possèdent la capacité de translater leur dynamique vers la luminance moyenne locale qu'ils reçoivent. Cette capacité est appelée compression adaptative logarithmique. Sur la figure 3.4, on observe que la sensibilité moyenne se translate et se centre sur la luminance locale moyenne, la dynamique du photorécepteur restant identique. Le comportement général sur la dynamique complète du photorécepteur se rapproche alors d'une loi logarithmique (pointillé sur la figure 3.4).

La compression réalisée par les photorécepteurs peut être modélisée par l'équation de Michaelis-Menten (Beaudot, 1994) adaptée pour des images en niveau de gris codées sur 8 bits (256 niveaux de gris) :

$$\begin{cases} r(p) = \frac{R(p)}{R(p) + R_0(p)} \cdot (255 + R_0(p)) \\ R_0(p) = \frac{V_0}{256} \cdot L(p) + (255 - V_0) \end{cases} \quad (35)$$

où p correspond à la position spatiale dans l'image, $r(p)$ est la luminance corrigée, $R(p)$ est la luminance d'entrée, $R_0(p)$ est un coefficient qui ajuste le gain du photorécepteur en fonction de la luminance moyenne $L(p)$. Le paramètre V_0 permet d'ajuster la force de la compression. Dans (Durette, 2005), l'auteur montre que V_0 peut être choisi dans un intervalle compris entre 160 et 250. Plus la valeur est faible, plus la compression est faible. Expérimentalement nous avons fixé V_0 à 230. Nous verrons dans les simulations que ce réglage permet une modélisation robuste de la zone de la bouche. La figure 3.5 présente les tracés de la réponse d'un photorécepteur en fonction de la valeur R_0 . On constate sur la figure 3.5 que les zones sombres de l'image, pour lesquelles R_0 sera faible, vont voire leur dynamique augmenter fortement. Au contraire quand R_0 sera grand, c'est-à-dire pour les zones dont la luminance moyenne sera grande, alors la dynamique locale sera peu compressée. Cette propriété sera particulièrement intéressante dans le cas d'image avec des zones très sombres et très claires. En particulier dans le cas des images de bouche, la zone interne de la bouche sera très souvent sombre à cause de la faible ouverture de la bouche ou de la direction de la lumière. Il faut souligner que la compression des photorécepteurs sera sans effet sur des images saturées. La figure 3.6 illustre l'effet de la compression adaptative des photorécepteurs sur une image de bouche. On observe une amélioration de la dynamique sur la zone interne de la bouche. La luminance interne est rehaussée et le contraste amélioré.

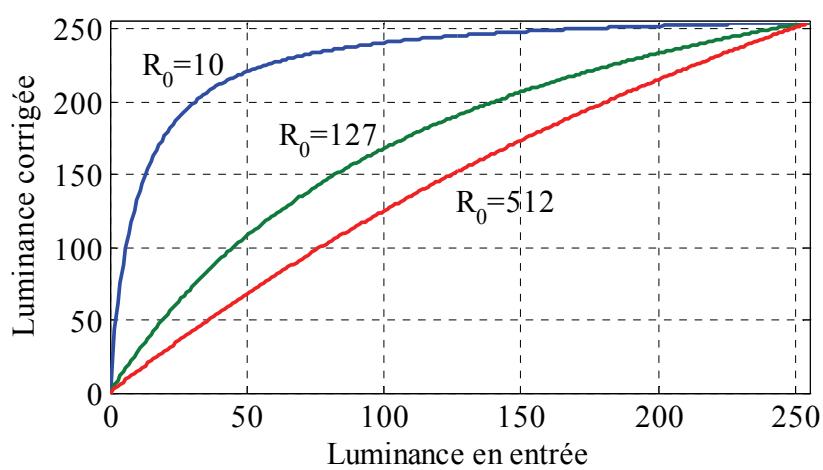


Figure 3.5 : Tracés des réponses d'un photorécepteur pour différentes valeurs de R_0 .

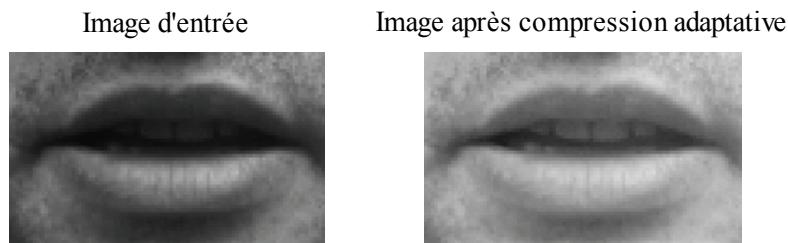


Figure 3.6 : Exemple de compression adaptative des photorécepteurs sur une image de bouche.

La compression logarithmique est le premier traitement de la rétine, l'information visuelle est par la suite transmise à la couche cellulaire suivante (cellules horizontales, bipolaires et amacrines) par l'intermédiaire de la *PLE*.

3.2.1.2 Couche Plexiforme Externe (*PLE*)

La *PLE* correspond à la zone de jonction entre les photorécepteurs, les cellules horizontales et les cellules bipolaires. Les photorécepteurs délivrent leur signal à la fois aux cellules horizontales et aux cellules bipolaires.

Cellules horizontales :

Les cellules horizontales connectent plusieurs photorécepteurs et sont connectées entre elles par l'intermédiaire de synapses. Chaque cellule horizontale collecte donc l'information provenant d'un nombre important de photorécepteurs. Une cellule horizontale sera donc influencée par les photorécepteurs de l'ensemble d'une zone de la rétine, on parle de champ récepteur. La conséquence de ces multiples connexions d'une cellule horizontale avec des photorécepteurs sera un lissage de l'information transmise par les photorécepteurs. Chaque cellule horizontale transmettra une information sur la luminance moyenne locale $L(p)$. Cette information permet, par rétroaction, d'effectuer la compression adaptative au niveau des photorécepteurs.

Cellules bipolaires :

Les cellules bipolaires relient les photorécepteurs et les cellules horizontales aux cellules ganglionnaires. Il existe 2 types de cellules bipolaires : les cellules bipolaires *ON* et les cellules bipolaires *OFF*. Pour ces cellules le champ récepteur est divisé en 2 zones : une

zone centrale et la zone périphérique. Les cellules bipolaires *ON* répondent lorsque le stimulus de la zone centrale est plus fort que le stimulus de la périphérie et inversement pour les bipolaires *OFF*. Les cellules bipolaires étant reliées aux photorécepteurs et aux cellules horizontales, le signal du champ récepteur central correspondra à la réponse des photorécepteurs et le signal du champ récepteur périphérique sera donné par les cellules horizontales. Ces opérations auront pour effet de supprimer la composante continue du signal visuel. Les cellules bipolaires seront alors sensibles aux variations locales de la luminance. En pratique, la réponse des cellules bipolaires *ON* est calculée en faisant la différence entre la sortie des photorécepteurs et la sortie des cellules horizontales avec mise à zéro des valeurs négatives, les cellules ne pouvant coder que des signaux positifs. Pour la réponse des cellules bipolaires *OFF*, on calcule la différence entre la sortie des cellules horizontales et des photorécepteurs et on ne garde que la partie positive. De cette manière, les parties positives et négatives du signal visuel sont codées, et il n'y a pas de pertes d'information. Les cellules bipolaires, *ON* et *OFF*, travaillent de manière complémentaire.

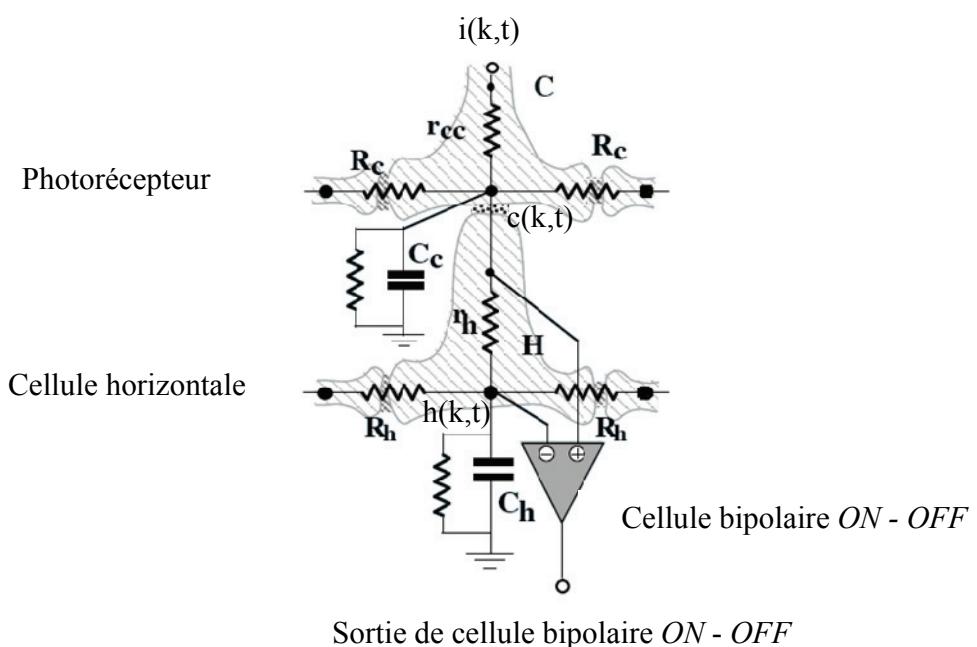


Figure 3.7 : Modélisation d'une triade synaptique (Hérault, 2001).

Modélisation de la PLE :

La *PLE* est la zone d'interconnexion entre le réseau de photorécepteurs, le réseau de cellules horizontales et les cellules bipolaires. Une jonction, appelée triade synaptique, permet l'interconnexion des photorécepteurs, des cellules horizontales et des cellules bipolaires. La *PLE* est constituée par un réseau de triade qu'on peut modéliser par le réseau électrique de la figure 3.7. La figure 3.7 montre que l'on a 2 niveaux de filtrage passe-bas spatio-temporel. Le premier niveau de filtrage passe-bas sera réalisé au niveau des photorécepteurs. Ce premier niveau de filtrage aura pour effet de limiter le bruit d'acquisition au niveau des photorécepteurs. Le second niveau de filtrage passe-bas permettra d'extraire la luminance locale $L(p)$. Enfin, les cellules bipolaires font l'extraction de la partie positive du signal lumineux pour les cellules *ON*, les cellules *OFF* codant la partie négative. Les signaux issus des cellules bipolaires *ON* et *OFF* sont combinés en faisant la différence des voies *ON* et *OFF*. Dans (Beaudot, 1994; Héault, 2001), les auteurs ont montré que l'on pouvait modéliser l'ensemble de la *PLE* par un réseau de triades synaptiques (figure 3.8).

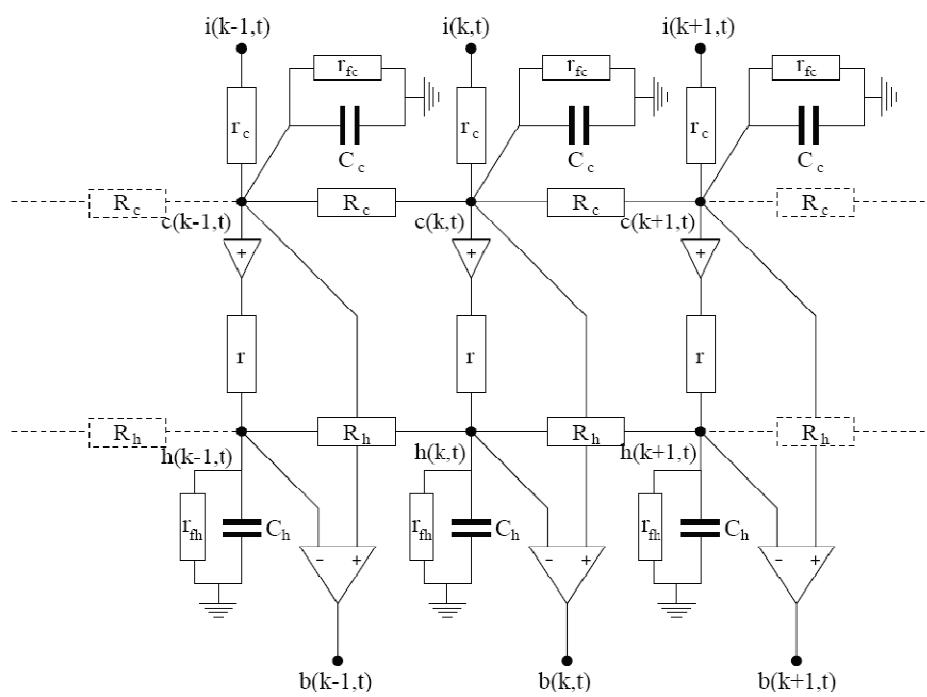


Figure 3.8 : Modèle électrique de la *PLE* (Beaudot, 1994).

Nous pouvons exprimer la fonction de transfert globale de la *PLE*. Pour le réseau de photorécepteurs, nous avons alors la fonction de transfert suivante (Beaudot, 1994 ; Hérault, 2001) :

$$F_c(fs, ft) = \frac{1}{1 + \beta_c + 2\alpha_c(1 - \cos(2\pi fs)) + j2\pi\tau_c ft} \quad (36)$$

Avec : $\alpha_c = r_c/R_c$, $\beta_c = r_c/r_{fc}$, $\tau_c = r_c C_c$

où r_c , R_c , r_{fc} et C_c correspondent aux résistances et aux capacités modélisant les photorécepteurs (figure 3.8). Pour le réseau des cellules horizontales, le transfert est analogue (Beaudot, 1994 ; Hérault, 2001) :

$$F_h(fs, ft) = \frac{1}{1 + \beta_h + 2\alpha_h(1 - \cos(2\pi fs)) + j2\pi\tau_h ft} \quad (37)$$

Avec : $\alpha_h = r_h/R_h$, $\beta_h = r_h/r_{fh}$, $\tau_h = r_h C_h$

où r_h , R_h , r_{fh} et C_h correspondent aux résistances et aux capacités modélisant les cellules horizontales (figure 3.8). Il faut ensuite combiner les sorties des cellules bipolaires *ON* et *OFF* :

$$G_{PLE}(fs, ft) = G_{BON}(fs, ft) - G_{BOFF}(fs, ft)$$

avec

$$\begin{cases} G_{BON}(fs, ft) = F_c(fs, ft)[1 - F_h(fs, ft)] \text{ si } F_c(fs, ft)[1 - F_h(fs, ft)] > 0 \\ G_{BON}(fs, ft) = 0 \text{ sinon} \end{cases} \quad (38)$$

et

$$\begin{cases} G_{BOFF}(fs, ft) = -F_c(fs, ft)[1 - F_h(fs, ft)] \text{ si } F_c(fs, ft)[1 - F_h(fs, ft)] < 0 \\ G_{BOFF}(fs, ft) = 0 \text{ sinon} \end{cases}$$

Au final, on obtient après simplification (Benoit, 2007) :

$$G_{PLE}(fs, ft) = F_c(fs, ft)[1 - F_h(fs, ft)]. \quad (39)$$

La figure 3.9 présente le tracé de la fonction de transfert globale de la *PLE*.

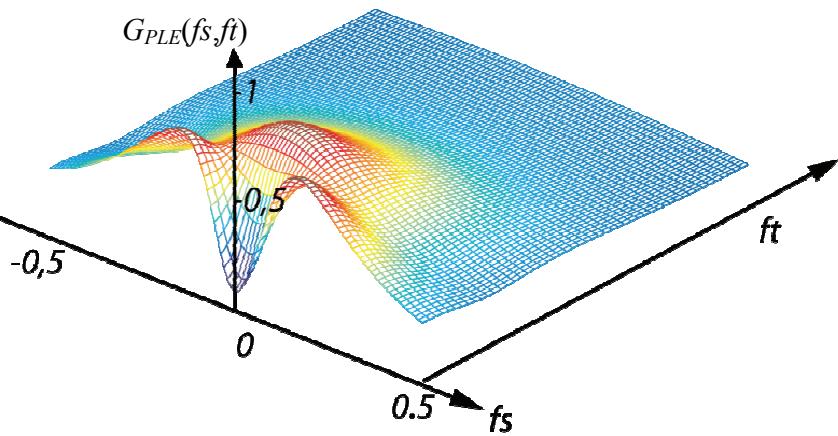


Figure 3.9 : Fonction de transfert $G_{PLE}(fs,ft)$

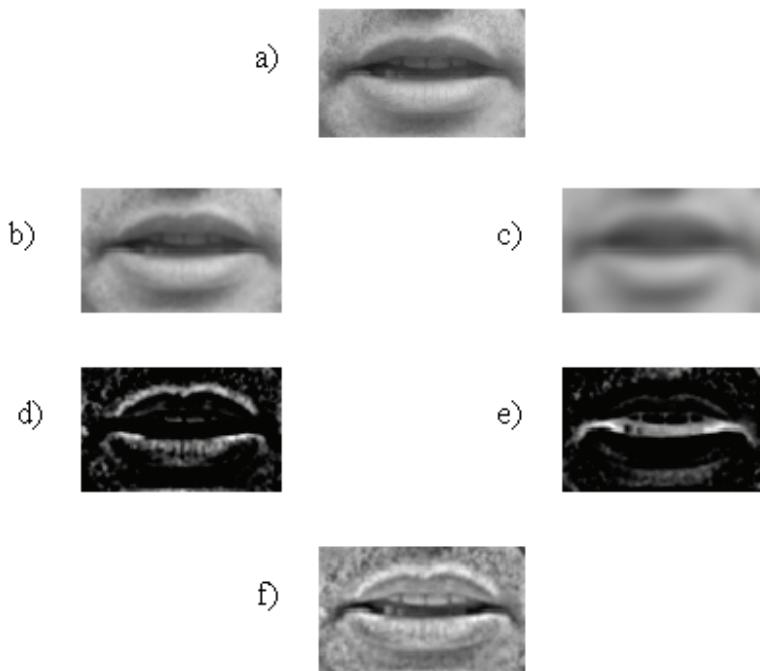


Figure 3.10 : Effet des différents réseaux de cellules de la *PLE*, a) image après compression adaptative, b) filtrage passe-bas des photorécepteurs, c) filtrage passe-bas des cellules horizontales, d) sortie des cellules bipolaires *ON*, e) sortie des cellules bipolaires *OFF*, f) sortie *ON-OFF*.

Dans notre cas, nous traitons des images de bouche statiques. Nous serons intéressés par l'effet de ce filtre pour $ft = 0$. La figure 3.9 montre que le *PLE* aura un effet passe bande (du type différence de gaussiennes) pour les fréquences temporelles faibles. Le réseau des

photorécepteurs réalise un premier filtrage passe-bas spatial, avec une fréquence de coupure haute élevée, destiné à atténuer le bruit entre les photorécepteurs (figure 3.10-b). Le réseau de cellules horizontales effectue un 2^{ième} filtrage passe-bas, avec une fréquence de coupure inférieure à celle des photorécepteurs. On obtient alors une estimation de la luminance locale moyenne (figure 3.10-c). Ensuite les cellules bipolaires *ON* effectuent la différence entre la réponse des photorécepteurs et celle des cellules horizontales tandis que les cellules bipolaires *OFF* effectuent l'inverse en ne gardant que l'information positive (figure 3.10-d et 3.10-e). En combinant les informations des sorties des cellules bipolaires *ON* et *OFF*, on obtient une image sur laquelle la luminance moyenne locale et le bruit haute fréquence ont été supprimés (figure 3.10-f), et sur laquelle les contours ont été renforcés.

Nous avons tracé les spectres en amplitude de l'image de bouche de la figure 3.10 avant la *PLE* et après filtrage par le modèle de *PLE* à la figure 3.11.

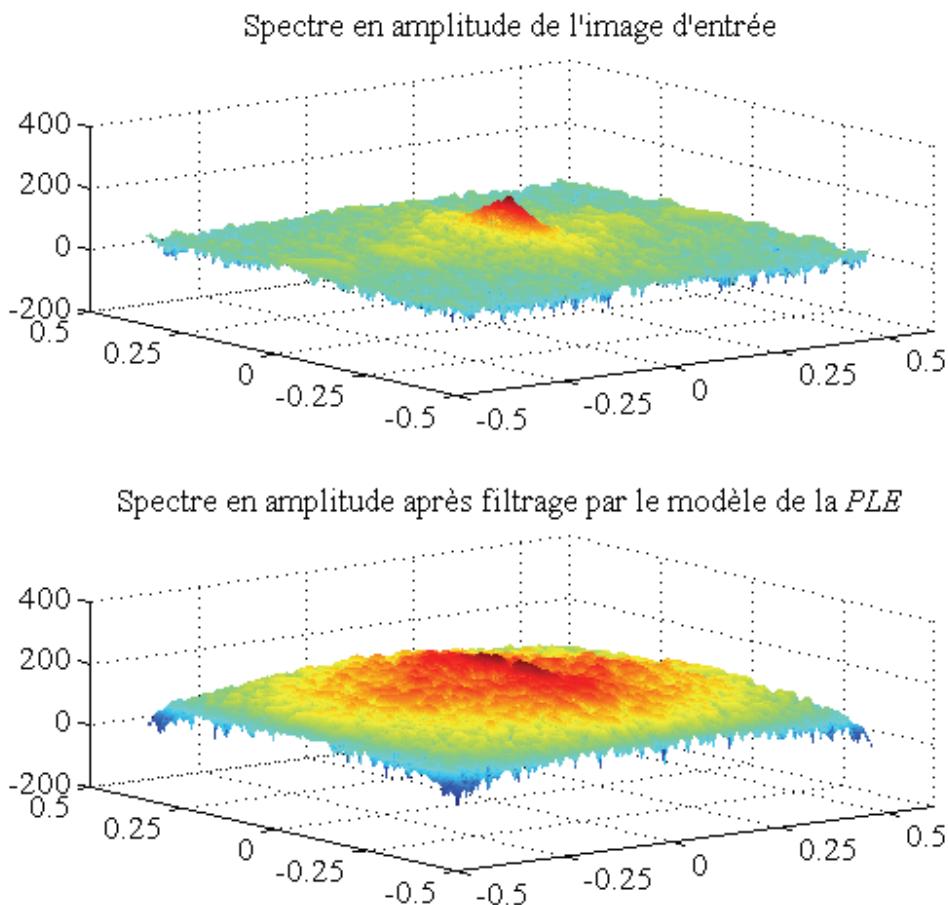


Figure 3.11 : Effet de blanchiment spectral de la *PLE*.

Nous constatons sur les tracés de la figure 3.11 que la *PLE* a un effet de blanchiment spectral sur les images. Les hautes fréquences de l'image, qui sont associées aux détails, en particulier les contours, sont rehaussées par le filtre modélisant la *PLE*. On distingue clairement sur le second tracé le rehaussement des hautes fréquences de l'image de bouche et une atténuation du mode principal du spectre de l'image filtrée, par rapport à l'image initiale où l'énergie du spectre est principalement concentrée sur les basses fréquences. Cela correspond bien au rehaussement des contours observé sur l'image de bouche à la figure 3.10-f.

3.2.1.3 Couche Plexiforme Interne (*PLI*)



Figure 3.12 : Signal visuel sur la voie Parvocellulaire en sortie de la rétine, a) image d'entrée, b) sortie *ON-OFF* sans compression adaptative des cellules ganglionnaire P, c) signal visuel *ON-OFF* transmis par la voie Parvocellulaire.

La *PLI* est la dernière couche de la rétine avant la transmission de l'information visuelle vers le cerveau par le nerf optique. L'information provenant de la *PLE* est transmise par les cellules bipolaires aux cellules ganglionnaires et amacrines par l'intermédiaire de la *PLI*. La sortie de la *PLI* est constituée par les axones des cellules ganglionnaires qui forment le nerf optique. Le nerf optique véhicule l'information visuelle jusqu'aux aires visuelles du cerveau. Il existe une grande diversité de cellules ganglionnaires et amacrines dont les comportements ne sont pas encore modélisés avec précision. En l'état actuel des connaissances, on distingue 3 canaux ou voies particulières résultant des interactions des différents types de cellules de la *PLI*: la voie Parvocellulaire (Parvo) qui est dédiée à la vision haute résolution et qui est sensible aux contrastes locaux dans l'image, la voie Magnocellulaire (Magno) dédiée aux informations de mouvements et la voie Koniocellulaire dont le rôle n'est pas clairement défini à l'heure actuelle. Du point de vue des images statiques, seul le canal Parvocellulaire, qui regroupe les traitements uniquement spatiaux de l'information visuelle, va nous intéresser. Le canal Parvocellulaire est constitué par les sorties des cellules ganglionnaires de type P qui agissent directement sur les sorties

des cellules bipolaires *ON* et *OFF*. Dans (Smirnakis, 1997), l'auteur a montré que ces cellules ganglionnaires adaptent leur réponse aux sorties des cellules bipolaires d'une manière analogue aux photorécepteurs, qui adaptent leur réponse à la luminance locale moyenne. Pour modéliser l'effet des cellules ganglionnaires de type P au niveau de la *PLI*, une compression adaptative est réalisée à la sortie des cellules bipolaires *ON* et *OFF* avec une loi de compression identique à celle utilisée pour les photorécepteurs. Ensuite seulement, les sorties *ON* et *OFF* sont combinées pour obtenir la sortie *ON-OFF*. La figure 3.12 illustre l'effet des cellules ganglionnaires de type P. La compression adaptative réalisée sur les sorties *ON* et *OFF* des cellules bipolaires permet de renforcer le contraste de l'image et d'augmenter la réponse sur les contours.

La figure 3.13 présente un résumé des différentes étapes de traitement du modèle de rétine pour des images statiques.

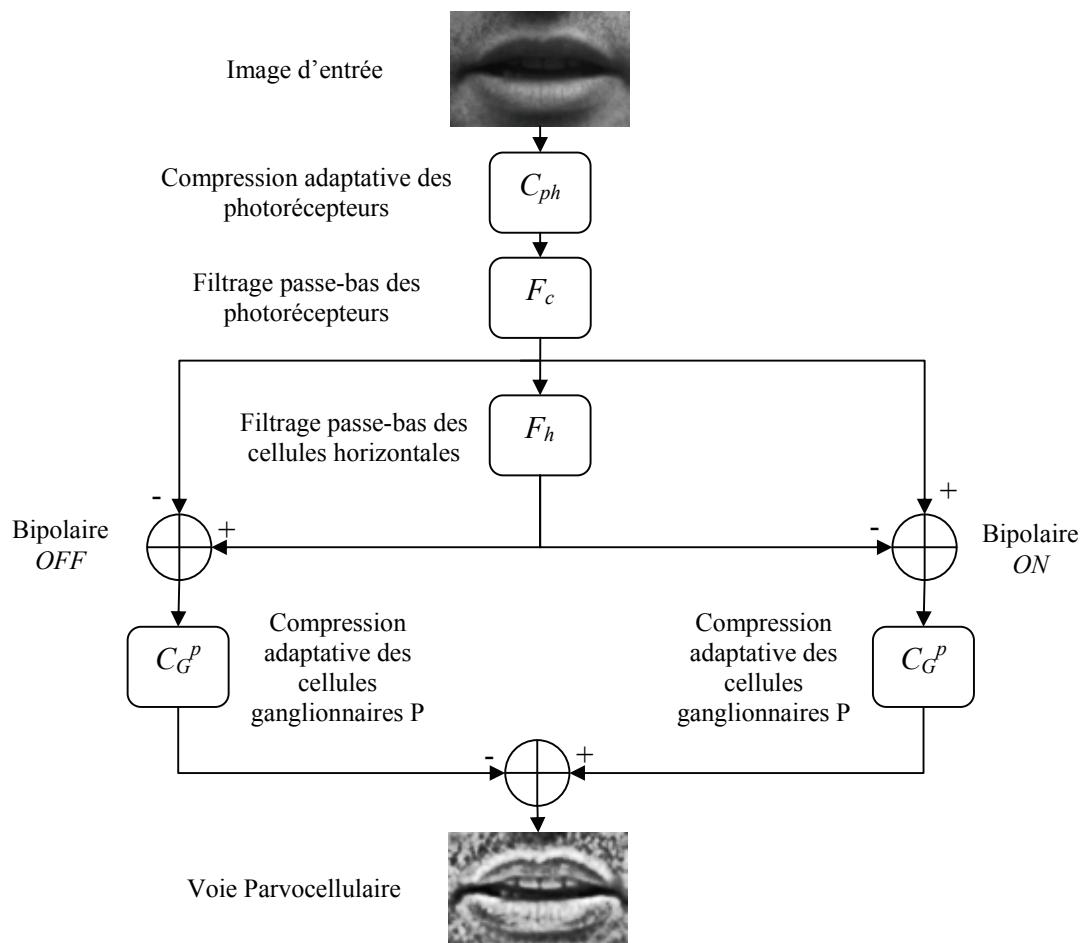


Figure 3.13 : Modélisation de la rétine pour l'analyse d'images statiques

3.2.2 Cortex V1

Les voies Parvocellulaire, Magnocellulaire et Koniocellulaire décrites précédemment sont véhiculées par les nerfs optiques gauche et droit en direction du cerveau. L'information visuelle transite tout d'abord par les corps genouillés latéraux (*CGL*). Le rôle principal des *CGL* est de relayer l'information visuelle. Cependant, dans (Sherman, 2002), l'auteur indique que les *CGL* assurent également un lien avec les aires motrices du cerveau. L'information visuelle en provenance des *CGL* est ensuite envoyée vers l'aire V1 du cortex occipital. L'aire V1 est la première étape du traitement de l'information visuelle par le cerveau. À l'heure actuelle, une trentaine d'aires ayant un rapport avec la vision ont été identifiées, mais seule l'aire V1 a été modélisée assez précisément. Les travaux de Hubel et Wiesel (Hubel, 1962), Blakemore et Campbell (Blakemore, 1969), De Valois (De Valois, 1982), Harvey et Doan (Harvey, 1990) ont montré que l'aire V1 procède à une analyse par bandes de fréquences et bandes d'orientations de l'information visuelle transmise par la rétine.

Pour modéliser la sensibilité aux orientations et aux bandes de fréquences des cellules de l'aire V1, plusieurs approches ont été proposées. Dans (Hawken, 1987), les auteurs utilisent des bancs de filtres de type différences de gaussiennes décalées. Les filtres de Gabor ont également été employés pour modéliser le comportement du cortex V1 (Marcelja, 1980 ; Daugman, 1988). L'hypothèse que le cortex V1 procède à une analyse en ondelettes de l'information visuelle a aussi été émise dans (Mallat, 1999).

Guyader, dans (Guyader, 2006), a proposé une modélisation plus précise du cortex V1 par l'introduction de filtres *Glop* ou filtres Gabor L^Og Polaire. La différence par rapport aux filtres de Gabor vient de l'ajout d'une contrainte de symétrie sur l'axe des fréquences en échelle logarithmique, ce qui se rapproche le plus de la manière dont l'information visuelle est décomposée dans l'aire V1. La contrainte de symétrie sur les filtres, en échelle logarithmique, est intéressante car elle permet de traiter les zooms et les changements d'échelles simplement. Nous verrons dans la suite qu'un zoom correspondra à une translation de l'énergie le long de l'axe des fréquences en échelle logarithmique. Cette modélisation présente l'avantage de donner une mesure des caractéristiques spectrales proche de celle du système biologique. Cette modélisation des images a été utilisée pour la

classification d'images statiques représentant des scènes naturelles (Guyader, 2006). Pour caractériser une image, la méthode consiste à échantillonner le spectre d'amplitude de l'image par une rosace de filtres *Glop* (figure 3.14).

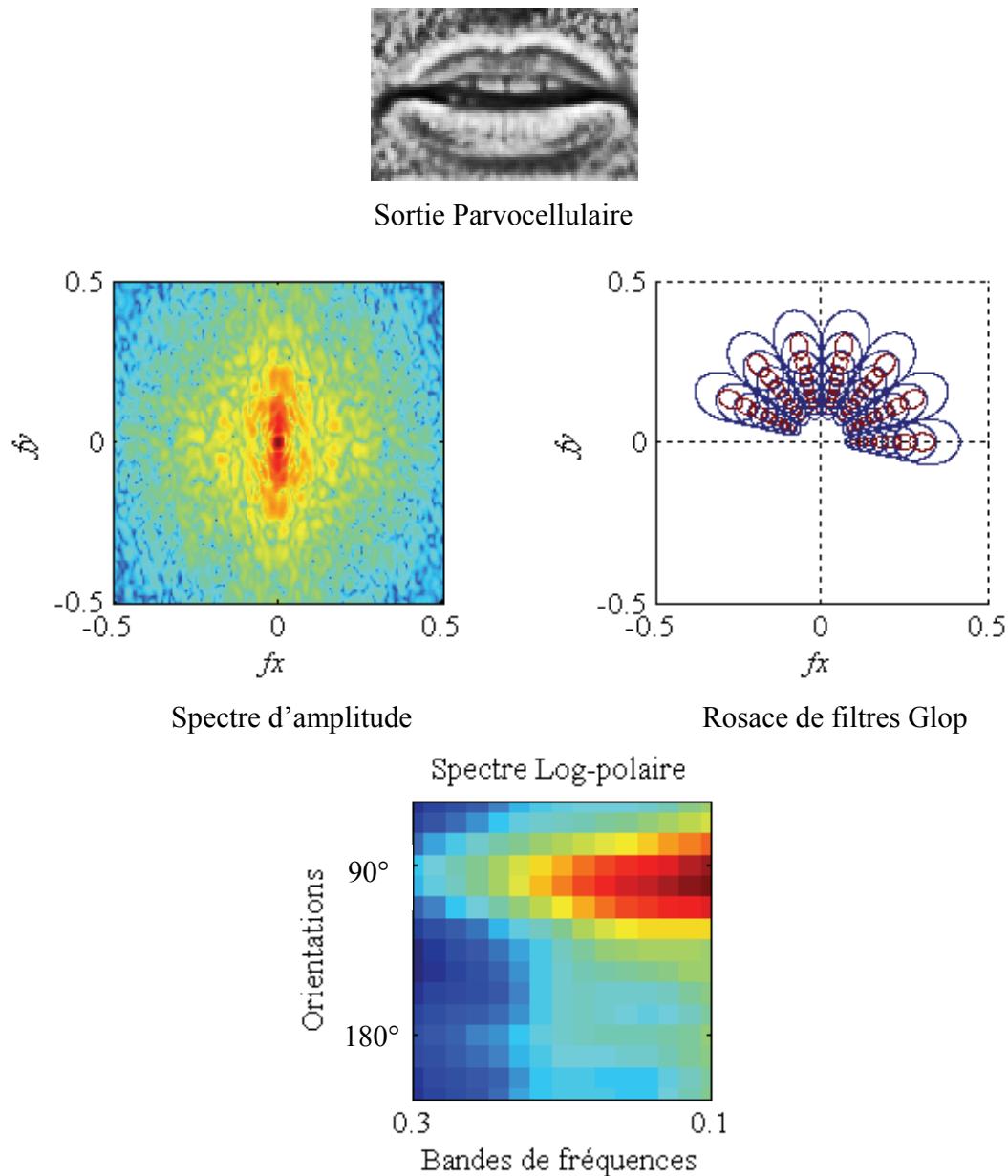


Figure 3.14 : Exemple de calcul d'un spectre Log-polaire. Le spectre d'amplitude de l'information visuelle, issue du modèle de rétine, est échantillonné par une rosace de filtres *Glop* pour obtenir le spectre Log-polaire.

Un filtre *Glop* est décrit par l'expression suivante :

$$G_{i,k}(f, \theta) = \frac{1}{2\pi\sigma^2} \frac{1}{f} \exp\left(-\frac{\ln\left(\frac{f}{f_k}\right)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta - \theta_i)^2}{2\sigma^2}\right) \quad (40)$$

Le filtre $G_{i,k}(f, \theta)$ à variables séparables est centré sur (f_k, θ_i) , f_k est la fréquence centrale sur l'axe des fréquences pour l'orientation θ_i . On remarque que le premier terme en f suit une loi log-normale (les filtres sont symétriques en log-fréquence, cf. figure 3.15) pour les fréquences, et que le terme en θ suit une loi gaussienne pour les orientations. La réponse du filtre nous donnera l'énergie moyenne dans une bande de fréquences et pour une orientation particulière.

La construction de la rosace de filtres Glop pour échantillonner le spectre d'amplitude d'une image est basée sur les travaux en neuroscience de (Blakemore, 1969 ; DeValois 1982). Ces travaux montrent que les cellules de l'aire V1 ont une largeur de bande à mi-hauteur comprise entre 1 et 2.5 octaves. En choisissant une largeur de bande à mi-hauteur de 1 octave ($f_{k+1}/f_k = 2$), nous obtenons alors des filtres sélectifs en fréquence. Avec une largeur de bande de 1 octave à mi-hauteur, on peut alors l'écart type σ :

$$\begin{aligned} 0.5 &= \exp\left(-\frac{\ln\left(\frac{f}{f_k}\right)^2}{2\sigma^2}\right) = \exp\left(-\frac{(\ln(2)/2)^2}{2\sigma^2}\right) \\ \Rightarrow \sigma &= \frac{\sqrt{2 \ln(2)}}{4} = 0.294 \end{aligned} \quad (41)$$

Etant donné la largeur σ des filtres, pour couvrir correctement le spectre des log-fréquences il faut alors choisir des fréquences centrales f_k décroissantes par octave. De plus, Guyader (Guyader, 2006) a montré que le nombre d'échantillons du spectre devait être impair. Soit f_{max} la fréquence centrale réduite maximale, les fréquences centrales des autres filtres sont obtenues par l'expression suivante : $f_k = f_{max}/1.5^i$.

La largeur transversale (en orientation) des filtres est adaptée en fonction du nombre d'orientations N_{angle} que l'on souhaite échantillonner avec comme contrainte que le recouvrement entre les filtres soit limité à la moitié de l'amplitude maximale.

Avec cette méthode, nous obtenons alors un spectre d'amplitude échantillonné par orientations et bandes de fréquences que l'on appelle spectre Log-polaire (figure 3.14). Sur le spectre de la figure 3.14 qui correspond au spectre échantillonné de l'image de bouche, on voit clairement que ce sont les orientations proches de 90° sur le spectre d'amplitude (l'horizontale sur l'image d'entrée) pour lesquelles la réponse des filtres Glop est maximale. Cela est cohérent avec l'image où la bouche est entrouverte. Le spectre Log-polaire offre une représentation réduite des caractéristiques de structure et de texture de l'image traitée par le modèle de rétine. Dans notre cas, nous avons fixé le nombre d'orientations et de bandes de fréquences à 15, dans les 2 cas, ce qui est assez proche du modèle biologique (Bullier, 2001). Les spectres Log-polaires auront alors une résolution de $15 \times 15 = 225$ pixels. Enfin, en choisissant $f_{max}=0.25$, les fréquences réduites au delà de 0.4, qui correspondent au bruit de mesure haute fréquence, seront atténuées (figure 3.15).

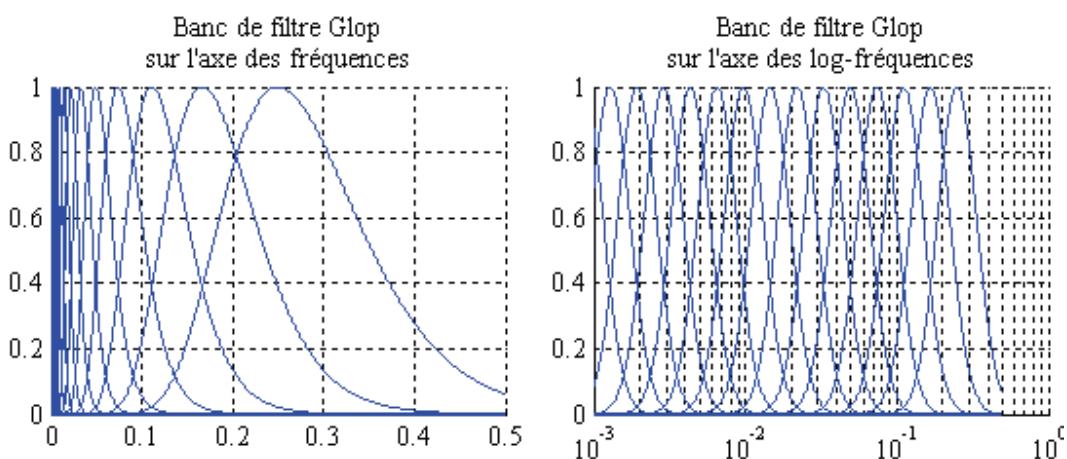


Figure 3.15 : Banc de filtres Glop en 1 dimension, à gauche nous avons tracé les profils des filtres sur l'axe des fréquences et à droite sur l'axe des log-fréquences. Les amplitudes des filtres ont été normalisées entre 0 et 1 pour améliorer la visualisation de la forme des filtres.

De plus, les spectres Log-polaires possèdent également des propriétés intéressantes relatives aux zooms et aux rotations. Soit une image $I(p)$ dont la transformée de Fourier est $S(f)$. Dans le cas d'un zoom a et d'une rotation R_ϕ , on alors : $I(aR_\phi) \Rightarrow 1/a^2 * S(1/a R_\phi f)$. Si

on se place en coordonnées polaires ($S(f) \Rightarrow S(e^v, \theta)$), on peut alors calculer la sortie d'un filtre Glop de la manière suivante :

$$\begin{aligned} C_{i,k} &= \iint_{v,\theta} S(e^{v-\ln(a)}, \theta - \theta_0) \cdot \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(v-v_k)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta-\theta_i)^2}{2\sigma^2}\right) dv d\theta \\ C_{i,k} &= \iint_{v,\theta} S(e^v, \theta) \cdot \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(v-v_k + \ln(a))^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta-\theta_i + \theta_0)^2}{2\sigma^2}\right) dv d\theta \end{aligned} \quad (42)$$

Il découle de (42) que :

- un décalage sur le spectre log-fréquence se traduit par une translation des échantillons (s'ils sont en nombre impair (Guyader, 2006)).
- Une rotation se traduit par une translation en coordonnées polaires.

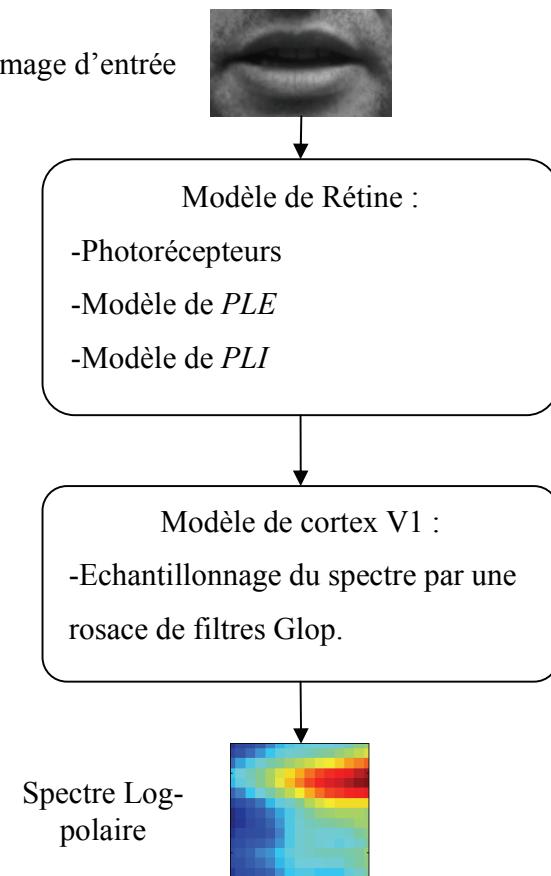


Figure 3.16 : Schéma global des traitements rétine+cortex V1

Dans cette section, nous avons décrit l'ensemble des traitements, inspirés du *SVH*, nous permettant d'aboutir à une description sous forme réduite de la structure d'une image. La figure 3.16 résume les étapes du processus de calcul du spectre Log-polaire d'une image.

3.3 Identification de l'état de la bouche

Dans ce chapitre, le but est de détecter si la bouche est ouverte ou fermée en utilisant l'information de luminance d'images statiques de bouche. Nous avons décrit, dans la section 3.2, un modèle du *SVH* qui permet de calculer un spectre Log-polaire. L'intérêt du spectre Log-polaire est de permettre la description de la structure de l'image traitée sous forme de vecteur. L'approche envisagée, dans ce chapitre, est d'utiliser les spectres Log-polaires comme signature pour identifier l'état de la bouche.

Dans (Benoit, 2005), l'auteur a utilisé une approche basée sur un modèle similaire pour détecter l'état dynamique de la bouche sur des séquences vidéos. L'hypothèse de Benoit était que l'énergie totale du spectre d'une bouche ouverte est plus grande que dans le cas d'une bouche fermée, à cause de la présence de plusieurs contours (dents, langue, ...). Des seuils adaptatifs sont appliqués sur l'énergie totale du spectre de la région de la bouche pour détecter l'état dynamique en fonction de l'évolution des maximums et des minimums d'énergie. La distribution spatiale du spectre n'est cependant pas prise en compte. Dans (Guyader, 2006), une approche supervisée utilisant les spectres Log-polaires a été utilisée pour classifier des images de scènes naturelles dans des catégories intuitives (images de montagnes, images de villes, images de plages, ...).

Notre idée est d'utiliser une approche supervisée basée sur les spectres Log-polaires pour identifier l'état de la bouche. Les spectres Log-polaires sont intéressants car ils permettent de décrire simplement, sous forme d'une matrice de dimension fixe, la structure d'une image. Cette description facilitera la comparaison entre des images avec des échelles et des résolutions différentes.

3.3.1 Base d'images de bouche

Dans un premier temps, nous avons construit une base d'images de bouche segmentées à l'aide de la méthode du chapitre 2. Pour constituer cette base, nous avons utilisé l'ensemble des images de la base comprenant 20 sujets dont 150 images avaient été extraites pour

l'étude des espaces couleur aux sections 1.2 et 2.3. La base d'images de bouche comprend 900 images de bouche de 20 sujets différents. Les images ont ensuite été classées manuellement en 2 catégories : bouche ouverte et bouche fermée. L'ensemble des bouches fermées comprend 230 images. L'ensemble des images de bouche ouverte comprend 670 images de bouche avec différents états d'ouverture. La bouche est considérée ouverte lorsque les dents, la langue ou la cavité buccale est visible (figure 3.17).



Figure 3.17 : Exemples d'images extraits de la base d'images de bouche segmentées à l'aide de la méthode du chapitre 2, sur la première ligne sont représentées des images de bouche ouverte et sur la seconde ligne des images de bouche fermée.

Les spectres Log-polaires ont été calculés pour les 900 images de la base. Nous avons considéré 15 bandes de fréquences et 15 orientations pour le calcul des spectres Log-polaires. Pour chaque image, on applique d'abord le modèle de rétine sur la luminance. Ensuite, on calcule la transformée de Fourier des images filtrées par le modèle de rétine. La transformée de Fourier suppose que l'on dispose d'un signal à support non-borné, ce qui n'est pas le cas des images de notre base. Si l'on calcule la transformée de Fourier directement en sortie du modèle de rétine, cela revient à travailler avec une image à support infini, apodisée par une fenêtre carrée. En fréquence, il y aura convolution entre la transformée de Fourier de l'image et la transformée de Fourier de la fenêtre carrée. Sur la première ligne de la Figure 3.18, on peut observer le spectre en amplitude de la sortie Parvocellulaire pour une image de bouche sans fenêtrage. On distingue clairement la distorsion en forme de croix, due à la convolution entre la transformée de Fourier de la sortie Parvocellulaire et la transformée de Fourier d'une fenêtre carrée. De plus, les

domaines spatial et fréquentiel étant duals, plus l'image sera petite, plus la distorsion sur la transformée de Fourier de l'image sera importante. Pour atténuer les effets induits par la taille finie de l'image, nous appliquons un fenêtrage de Hanning avant le calcul de la transformée de Fourier de l'image. On voit sur la seconde ligne de la figure 3.18 que le fenêtrage de Hanning atténue les distorsions.

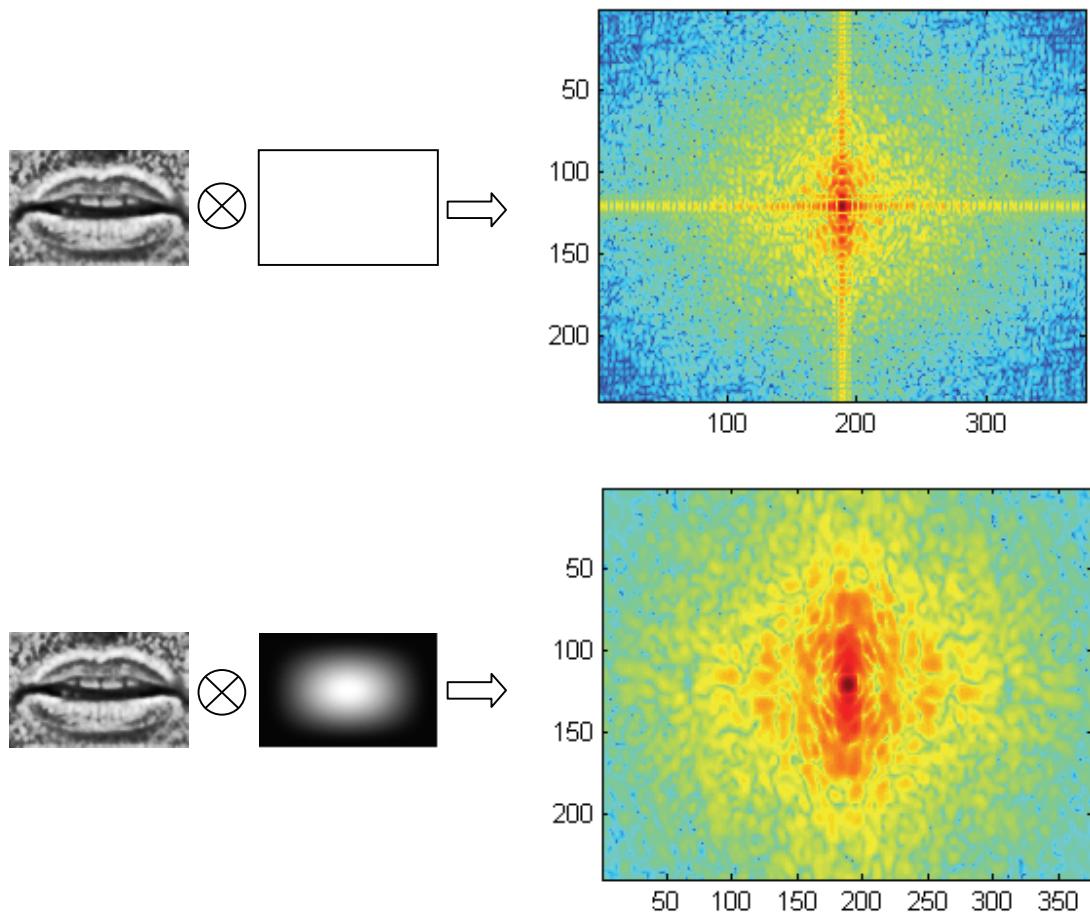


Figure 3.18 : Effet du fenêtrage de Hanning sur la transformée de Fourier de la sortie Parvocellulaire de la rétine pour une image de bouche.

Cependant ce fenêtrage ne sera pas sans effet sur la transformée de Fourier de la sortie Parvocellulaire. La figure 3.19 donne la forme de la fenêtre de Hanning en 1 dimension ainsi que l'amplitude de la transformée de Fourier de la fenêtre de Hanning. La convolution dans le domaine fréquentiel entre la transformée de Fourier de la sortie Parvocellulaire et la transformée de Fourier de la fenêtre de Hanning engendrera un moyennage du spectre de la sortie Parvocellulaire. Plus l'image sera petite, plus cet effet sera important, du fait de la

dualité du domaine spatial et du domaine fréquentiel, avec comme conséquence la perte des informations de détails de l'image. Dans notre cas nous avons appliqué un fenêtrage de Hanning.

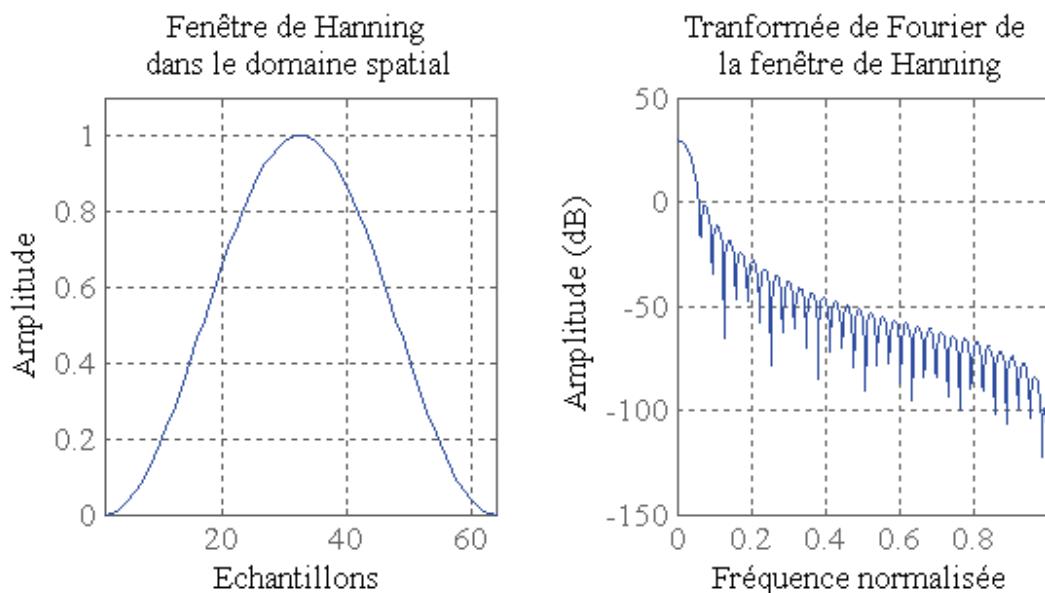


Figure 3.19 : Transformée de Fourier de la fenêtre de Hanning

3.3.2 Analyse en Composantes Principales (*ACP*)

Après avoir calculé les spectres Log-polaires pour les images de la base de bouche, nous avons procédé à une analyse en composantes principales (*ACP*) sur l'ensemble des spectres, concaténés pour former des vecteurs colonnes de 225 valeurs. Nous disposons alors d'un ensemble de données de dimension 225x900. L'analyse en composantes principales permet d'extraire les modes de variations principaux, ainsi que la variance sur chaque mode principal, sur un échantillon de données. La figure 3.20 présente la projection des spectres Log-polaires des images de notre base sur les 2 modes principaux qui représentent 85% de la variance totale des données. Ce type de projection, sur les 2 premières composantes principales, permet de voir s'il se dégage des tendances sur l'ensemble des données. Le nuage de points bleus correspond aux bouches fermées et le nuage de points rouges correspond aux bouches ouvertes.

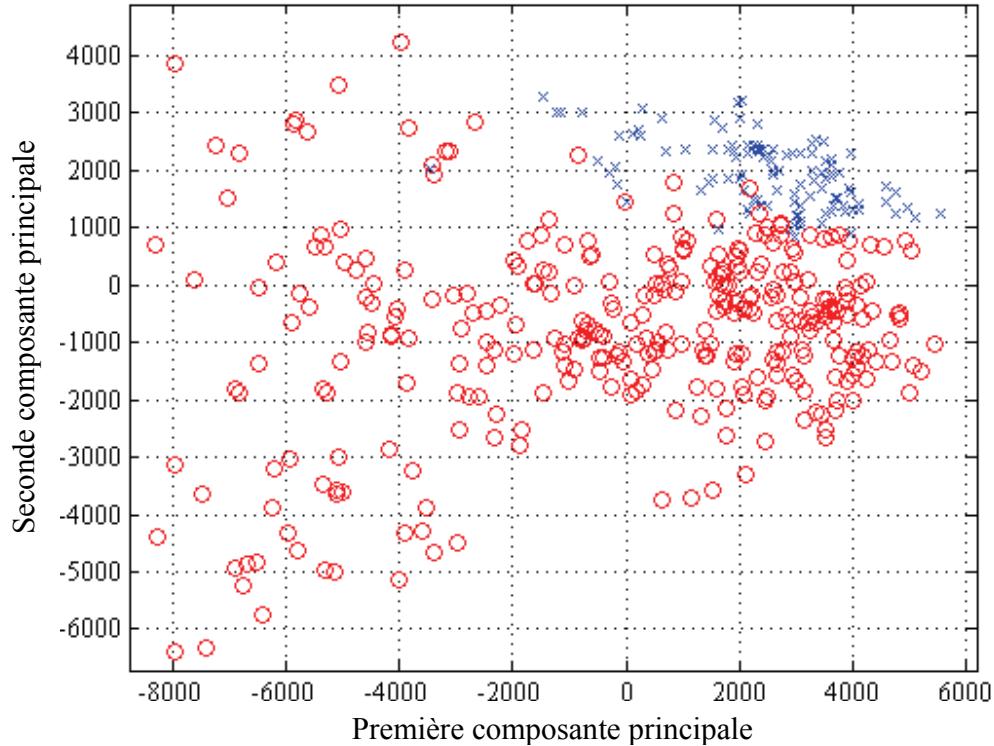


Figure 3.20 : Projection des spectres Log-polaires des images de bouche sur les 2 premières composantes principales données par ACP. Le nuage de points bleus correspond aux bouches fermées, le nuage de points rouges correspond aux bouches ouvertes.

On observe sur la figure 3.20 que les projections mettent en évidence une organisation des données. La topologie des projections des spectres Log-polaires des 2 catégories d’images de bouche forme 2 ensembles de points. Le nuage de point représentant les bouches fermées est relativement compact tandis que le nuage de points de la catégorie bouche ouverte est beaucoup plus dilaté. Intuitivement, on peut interpréter ce phénomène par les différences d’ouverture des bouches de notre base. Les projections sur les 2 premières composantes principales des données montrent que la description des images de bouche par les spectres Log-polaires permet d’obtenir une organisation cohérente des données. C'est-à-dire que nous obtenons 2 ensembles de points convexes avec un léger chevauchement. Cette analyse, sur les 2 premières composantes principales, nous permet de valider la méthode de modélisation des images de bouche. Par la suite, nous serons amenés à prendre

en compte un pourcentage de la variance totale plus grand (95%) pour la reconnaissance de l'état ouvert/fermé de la bouche. Cela impliquera de considérer plus de 2 modes principaux.

3.3.3 Classification supervisée des images de bouche

Nous avons vu dans la section précédente que les modèles de rétine et de cortex V1 permettent d'obtenir une description cohérente des bouches ouvertes et des bouches fermées (figure 3.20). De plus, les propriétés du modèle de rétine, compression adaptative de la luminance (photorécepteurs), suppression de la composante continue de l'information visuelle (cellules bipolaires *ON* et *OFF*), entraîne une normalisation de l'information visuelle sur la sortie Parvocellulaire. Plusieurs approches peuvent être envisagées pour classifier les images de bouches. Etant donné que la chaîne de traitement, composée du modèle de rétine et du modèle de cortex suivi par une *ACP* sur les spectres Log-polaires, permet de décrire les images par des vecteurs de taille fixe, l'approche classique est de construire un réseau de neurones de type perceptron multicouche pour effectuer la catégorisation des images de bouches. Les réseaux du type perceptron multicouche permettent de placer des frontières de décision linéaires dans un espace de données quelconque. Par ailleurs, il a été montré qu'un réseau à 3 couches, avec suffisamment de neurones sur les couches cachées, est un estimateur universel. La principale difficulté rencontrée, lors de l'utilisation des perceptrons multicouches, réside dans la recherche de la topologie adéquate du réseau pour modéliser l'ensemble de données dont on dispose. À l'heure actuelle, il n'existe pas vraiment de formalisme pour définir la topologie d'un perceptron multicouche. La construction d'un réseau de neurones de ce type est généralement intuitive et fonction de la quantité d'échantillons dont on dispose. La plupart du temps, les vecteurs en entrée du réseau sont de dimension supérieure à 3 ce qui rend l'observation de la topologie des données impossible. La meilleure architecture du réseau est souvent obtenue par des essais successifs. En particulier, dans le cas d'un problème de classification non-linéaire, il est difficile de construire un perceptron multicouche.

Pour l'identification de l'état ouvert/fermé de la bouche, nous avons privilégié une approche utilisant un réseau à machines à vecteurs de support ou Support Vector Machine (*SVM*). Les réseaux *SVM* sont une classe d'algorithmes d'apprentissage supervisés développés à l'origine pour le problème de la séparation non-linéaire à 2 classes. Les *SVM*

reposent sur 2 idées fondamentales. Tout d'abord, la recherche de la frontière de décision entre les 2 classes se fera de la manière suivante : on cherchera la frontière de décision pour laquelle la distance entre les échantillons des 2 classes et la frontière est maximale, ces échantillons sont appelés vecteurs de support (figure 3.21).

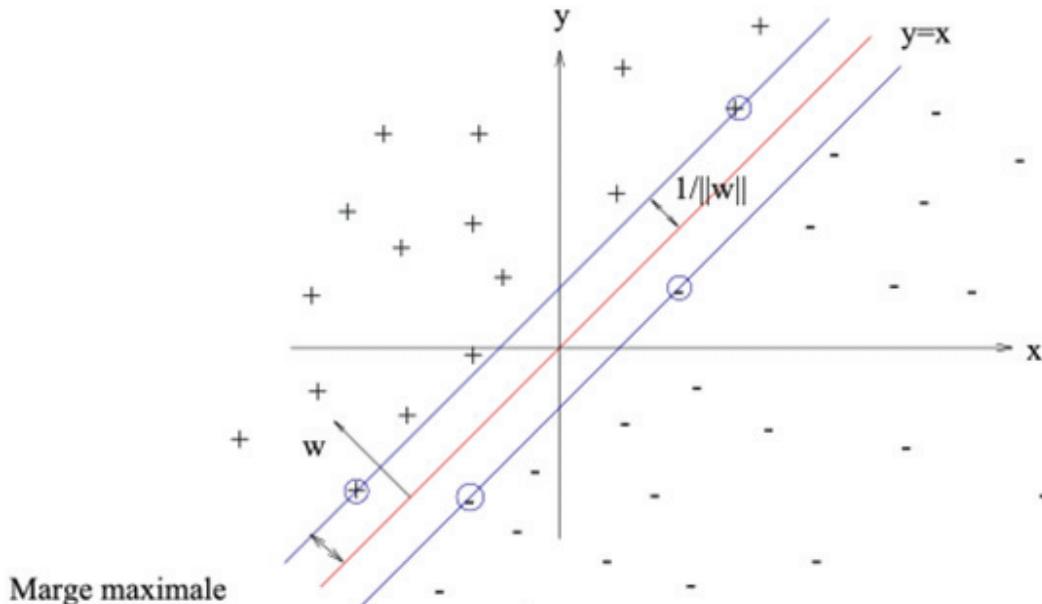


Figure 3.21 : Hyperplan optimal dans le cas d'un problème de séparation linéaire à 2 classes. En rouge est tracée la frontière de décision, les échantillons entourés en bleu sont les vecteurs de support. (http://fr.wikipedia.org/wiki/Machine_a_vecteurs_de_support).

Deuxièmement, pour les problèmes de séparation non-linéaire, l'idée avec les *SVM* est d'appliquer aux données une transformation vers un espace dans lequel il est possible de trouver une frontière de séparation linéaire. Pour réaliser cette transformation, il faut définir un noyau. La plupart du temps, on utilise comme noyau une fonction à base radiale (cf. technique dite du «kernel trick») (Aizerman, 1964)). Cette technique permet d'utiliser un classifieur linéaire pour résoudre un problème non-linéaire. La discrimination linéaire dans le nouvel espace revient alors à faire une discrimination non-linéaire dans l'espace d'origine. Dans notre cas, les projections de la figure 3.20 montrent que les ensembles de bouches ouvertes et fermées forment 2 nuages de points convexes avec un léger recouvrement. Il n'y a, toutefois, pas de garantie que les données soient linéairement séparables si plus de dimensions sont prises en compte. Dans la suite, nos vecteurs de

données, après *ACP*, seront de dimension égale à 10. Le noyau qui a permis d'obtenir la meilleure modélisation des données par le réseau *SVM* est un noyau de type gaussien.

Les lecteurs intéressés par le formalisme et l'optimisation des réseaux *SVM* pour les problèmes de classification non-linéaire pourront se reporter aux travaux de (Cortes, 1995) pour une description détaillée.

3.4 Résultats expérimentaux

Pour valider notre chaîne de traitement des images de bouche nous avons procédé à plusieurs séries de tests de classification. Tout d'abord, nous avons testé les résultats de la classification dans le cas où le réseau *SVM* est entraîné avec les données des images traitées avec l'ensemble de la chaîne de traitement modèle de rétine et modèle de cortex V1. Les spectres Log-polaires ont été calculés pour l'ensemble des 900 images de notre base d'apprentissage. Une *ACP* est ensuite effectuée, on garde 95% de la variance totale des données. On réduit ainsi la dimension de l'espace des données aux 10 premières composantes principales. Nous disposons de 900 vecteurs de dimension 10, et des états fixés manuellement pour entraîner le réseau *SVM*. Pour tester la performance du classifieur, nous avons d'abord calculé le taux de bonnes classifications lorsque tous les sujets sont inclus dans l'entraînement du réseau *SVM* (table 3.1). La table 3.1 présentent également le détail des résultats de classification sous forme de matrice de confusion.

Apprentissage global 98.2% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	656	14
	Bouche fermée	2	228

Table 3.1 : Résultats de classification lorsque les images de bouche sont traitées par l'ensemble de la chaîne modèle de rétine et modèle de cortex V1 lorsque tous les sujets sont inclus dans l'entraînement du réseau *SVM*.

Ensuite, nous avons de nouveau calculé le taux de classification sur les 900 images en utilisant pour chaque sujet un classifieur particulier. Pour chaque sujet de la base, nous avons exclu de l'apprentissage du classifieur particulier l'ensemble des images de ce sujet (test du « Leave-one-out ») et nous avons catégorisé ces images exclues avec ce classifieur.

Les images de chaque sujet de la base sont traitées avec un classifieur n'ayant aucune connaissance a priori sur le sujet. Nous donnons dans la table 3.2 le taux de classification pour l'ensemble des images de la base (les 900 images) ainsi que le détail des résultats de classification sous forme de matrice de confusion.

Test du "leave-one-out" 96.4% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	646	24
	Bouche fermée	8	222

Table 3.2 : Résultats de classification des images de bouche pour le test du « leave-one-out ».

Lorsque toutes les images de la base d'apprentissage sont utilisées pour entraîner le réseau *SVM*, on obtient un taux global de classifications correctes de 98.4% (table 3.1). La table 3.1 montre qu'il y a proportionnellement plus d'erreurs d'identification avec les bouches ouvertes. On constate également que le taux de faux positifs est plus élevé que le taux de faux négatifs pour les bouches fermées. Le taux de classification dans le cas du test du « leave-one-out » est, sans surprise, inférieur au cas où toutes les images sont incluses dans l'apprentissage du classifieur. Il reste cependant supérieur à 95 %, ce qui dénote une bonne capacité de généralisation des différents modèles. Dans le détail, la table 3.2 montre que, les performances se dégradent le plus pour la reconnaissance des bouches fermées. Pour les 2 catégories d'images de bouche le taux de bonnes classifications est cependant supérieur à 96 %. En ce qui concerne les mauvaises classifications, la figure 3.22 présente les images manuellement classées comme fermées qui ont été identifiées comme ouvertes par le classifieur. De la même manière, nous présentons à la figure 3.23 les images manuellement classées comme ouvertes et qui ont été identifiées comme fermées par le classifieur durant nos tests.



Figure 3.22 : Images de bouche identifiées subjectivement comme fermées et reconnues comme des bouches ouvertes par le classifieur.



Figure 3.23 : Images de bouche identifiées subjectivement comme ouvertes et reconnues comme des bouches fermées par le classifieur.

Globalement, les figures 3.22 et 3.23 montrent que les images classifiées de manière erronée correspondent à des bouches faiblement ouverte. La classification subjective peut être mise en cause également. Sur les 2 images de droites de la figure 3.22, les bouches peuvent être considérées ouvertes. Nous n'avons pas observé de classification aberrante, comme le cas d'une bouche largement ouverte et identifiée comme fermée. Sur les images mal classées, la taille de l'ouverture de la bouche est en moyenne de 6 pixels avec une résolution moyenne des images de 54x107.

Par la suite, nous avons testé l'influence du modèle de rétine et du modèle de cortex V1 sur les résultats de reconnaissance d'images de bouche. Pour tester l'influence du modèle de rétine sur la performance de la reconnaissance de l'état ouvert/fermé de la bouche, nous avons répété les simulations précédentes en remplaçant le modèle de rétine (compression des photorécepteurs, *PLE* et *PLI*) par une simple correction de l'illumination. La dynamique des images est centrée sur zéro et elle est normalisée par son écart type. L'image normalisée est ensuite traitée par le modèle de cortex V1. Nous avons répété les mêmes tests que dans le cas où l'ensemble de la chaîne de traitement est utilisée. Un réseau *SVM* est entraîné pour classifier les images de bouches. La table 3.3 donne les résultats de classification quand toutes les images de la base sont utilisées pour entraîner le classifieur.

Apprentissage global : 96.9% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	650	20
	Bouche fermée	8	222

Table 3.3 : Résultats de classifications lorsque le modèle de rétine est remplacé par une simple correction de luminance (moyenne nulle et variance unitaire) pour l'apprentissage global.

La table 3.4 donne les résultats de classification lorsque le modèle de rétine est remplacé par une simple correction de luminance (moyenne nulle et variance unitaire) pour le test du « leave-one-out ».

Test du "leave-one-out"		Résultats de classification	
86.7% de classification correcte		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	592	78
	Bouche fermée	39	183

Table 3.4 : Résultats de classifications lorsque le modèle de rétine est remplacé par une simple correction de luminance (moyenne nulle et variance unitaire) pour le test du « leave-one-out ».

On constate une baisse de performance de la classification des images de bouche avec la normalisation simple de la luminance par rapport au cas où le modèle de rétine est employé. En particulier, lorsqu'aucune image du sujet testé n'est utilisée dans l'apprentissage du classifieur (test du « leave-one-out »), on note une chute de 10% du taux de classification. Les tables 3.3 et 3.4 donnent également le détail des résultats de classification par catégorie d'images de bouche pour les 2 types d'entraînement du classifieur. Globalement, sans le modèle de rétine, lorsque toutes les images d'un sujet sont retirées de l'apprentissage du classifieur, la capacité de généralisation est bien inférieure au cas où la rétine est utilisée pour traiter la luminance des images. Le traitement par le modèle de la rétine permet une meilleure normalisation de l'information de luminance que la simple normalisation avec une moyenne nulle et une variance unitaire.

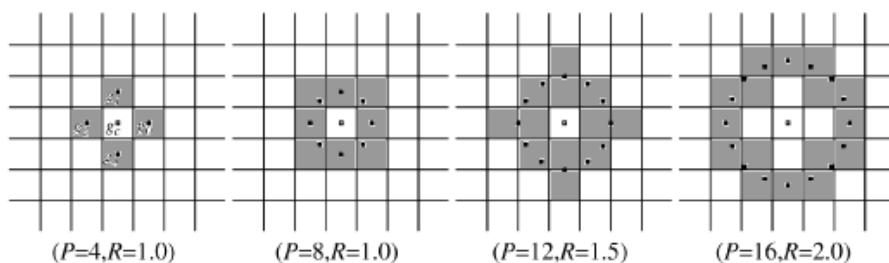


Figure 3.24 : Voisinages circulaires (Ojala, 2002)

Nous avons également testé l'intérêt du modèle de cortex V1 en le remplaçant par une description du type patron binaire local multi-résolution (Ojala, 2002) (Multiresolution

Local Binary Patterns). Cette méthode est couramment utilisée dans les problèmes de classification de textures. Les patrons binaires locaux, ou *LBP* (Local Binary Pattern), sont définis sur des voisinages circulaires (figure 3.24).

Pour chaque pixel de niveau de gris g_c d'une image, on extrait la distribution de niveaux de gris du voisinage circulaire $T_{g_c} = t(g_c, g_0, \dots, g_{P-1})$, où P est le nombre de points du voisinage, et les points $\{g_0, \dots, g_{P-1}\}$ sont les niveaux de gris des P points uniformément répartis sur le cercle de rayon R . Pour extraire les variations locales de luminance, le niveau de gris central g_c du voisinage est retranché aux niveaux des autres pixels du voisinage. Un seuil est ensuite appliqué :

$$T_{g_c} \approx t(s(g_0 - g_c), \dots, s(g_{P-1} - g_c))$$

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (43)$$

Le patron binaire correspond à une opération de seuillage des niveaux de gris des points du voisinage. Pour obtenir une description invariante aux rotations, (Ojala, 2002) calcule le coefficient $LBP_{R,P}(g_c)$ de la manière suivante pour tous les pixels de l'image :

$$LBP_{R,P}(g_c) = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{si } U(LBP_{R,P}(g_c)) \leq 2 \\ P+1, & \text{sinon} \end{cases}$$

$$U(LBP_{R,P}(g_c)) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \quad (44)$$

Cette description des variations locales de luminance permet, pour un voisinage donné (R et P fixés), de calculer un histogramme, par accumulation des coefficients $LBP_{R,P}(g_c)$, sur l'ensemble des pixels de l'image (Ojala, 2002). La dimension de l'histogramme sera, d'après (44), de $P+1$. Les images seront décrites par des vecteurs de dimensions fixes. En combinant les informations de voisinages de tailles différentes, la description peut aussi être robuste aux variations de résolution (Ojala, 2002).

Pour évaluer l'intérêt de la description des images de bouche fournie par le modèle de cortex V1, nous avons testé l'entraînement du réseau *SVM* en utilisant la description des images, traitées par le modèle de la rétine, fournie par une méthode *LBP* avec 3 voisinages ($\{R=1, P=8\}, \{R=2, P=16\}, \{R=3, P=32\}$). Pour chaque image, les histogrammes des coefficients $LBP_{R,P}(g_c)$ pour les 3 voisinages sont concaténés et forment les entrées du réseau *SVM*. Comme pour le test de l'ensemble de notre chaîne de traitement, nous avons calculé les taux de classification des images de bouche lorsque le classifieur est entraîné avec les données de toutes les images de la base et pour le test du « leave-one-out ». Les résultats sont donnés dans les tables 3.5 et 3.6. Les résultats des tables 3.5 et 3.6 montrent que la performance de la classification des images est très inférieure à la performance obtenue avec l'ensemble de notre chaîne de traitement des images de bouche. En particulier, le taux de classification lors du test du leave-one-out est très inférieur au taux enregistré lors du test de notre chaîne de traitement. On passe de 96.4% à 85.7%, avec une importante chute du taux de reconnaissance des images de bouche fermée. La description des images de bouche fournie par le modèle de cortex V1 semble donc pertinente pour le problème d'identification de l'état de la bouche.

Apprentissage global : 97.56% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	655	15
	Bouche fermée	8	222

Table 3.5 : Résultats de classification lorsque le modèle de cortex V1 est remplacé par la méthode *LBP*.

Test du "leave-one-out" 85.7% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	616	54
	Bouche fermée	74	156

Table 3.6 : Résultats de classification lorsque lorsque le modèle de cortex V1 est remplacé par la méthode *LBP* pour le test du « leave-one-out ».

Enfin, pour valider notre chaîne de traitement et notre classifieur, nous avons testé notre méthode d'identification sur des images de bouches tirées de la base AR (Martinez 1998).

Cette base est composée d'images provenant de 126 sujets (voir section 1.6.1) avec différentes poses, conditions d'éclairage et portant divers accessoires. Nous avons extrait 473 images de bouches fermées et 507 images de bouches ouvertes. Pour entraîner le classifieur, nous avons utilisé uniquement les images de notre base traitée avec l'ensemble de notre chaîne de traitement, qui comprend le modèle de rétine, suivi du modèle de cortex V1. Aucune image de la base AR n'a été utilisée dans l'entraînement du classifieur. Le taux de bonnes classifications est de 98.5%. La table 3.7 donne les résultats par catégorie d'image de bouche.

Test avec les images de la base AR : 98.5% de classification correcte		Résultats de classification	
		Bouche ouverte	Bouche fermée
Vérité-terrain	Bouche ouverte	503	4
	Bouche fermée	10	463

Table 3.7 : Taux de classification par catégorie d'images de bouche pour la base AR.

3.5 Bilan

Dans ce chapitre, nous avons présenté une méthode supervisée pour identifier l'état (ouvert ou fermé) de la bouche sur des images statiques. Cette méthode repose sur une chaîne de traitement inspirée par le système visuel humain. Les résultats de nos simulations montrent la pertinence du modèle de rétine pour traiter l'information de luminance des images de bouche et du modèle de cortex V1 pour construire une description des images. Le taux de reconnaissance de notre classifieur d'images de bouche reste supérieur à 95% même lorsque celui-ci est testé sur des sujets inconnus. L'ensemble modèle de rétine et modèle de cortex V1 produit une information robuste aux variations de luminance et aux variations de morphologie des sujets. De plus, même dans le cas de classification erronée, il n'a jamais été constaté de cas aberrant où une bouche grande ouverte aurait été reconnue comme une bouche fermée. Les erreurs de classification sont survenues pour des images avec une ouverture de la bouche faible, de l'ordre de 6 pixels avec des images d'une résolution moyenne de 54x107 pixels. Pour les besoins de nos travaux, les modèles de rétine et de cortex V1 ont été implémentés dans Matlab. Pour une image de bouche d'une résolution de 152*54 pixels, le temps de calcul pour détecter l'état de la bouche est d'environ 1s sur une

machine équipée d'un processeur double cœur, avec une fréquence d'horloge de 2.33 GHz et de 2 Go de mémoire RAM.

Chapitre 4. Etude de la modalité infrarouge pour la segmentation des lèvres

4.1 Introduction

Historiquement, les techniques d'analyse faciale telles que la reconnaissance de visage ou la segmentation d'indices visuels, ont été développées à partir d'images acquises dans le domaine visible, les systèmes d'acquisition d'images dans le domaine visible étant disponibles et peu chers. De plus, la plupart des applications en analyse faciale sont liées à l'interaction entre humains. Il est naturel de privilégier le domaine visible du spectre pour développer des techniques d'analyse faciale. L'exploitation de nouvelles modalités dans les problèmes d'analyse faciale est devenue un axe de développement actif depuis quelques années. La modalité infrarouge, en particulier, a été utilisée dans des applications de reconnaissance faciale (Yoshitomi, 1997; Socolinsky, 2003). Dans les chapitres précédents, nous avons abordé le problème de la segmentation des lèvres en nous basant également sur des images acquises dans le domaine visible du spectre électromagnétique. Nous avons pu voir que les conditions d'illumination influençaient la robustesse des algorithmes d'analyse labiale. Suivant la direction de la source lumineuse, il peut apparaître des ombres et des réflexions spéculaires sur les lèvres ou d'autres éléments de la zone de la bouche. Ces ombres et réflexions peuvent provoquer l'apparition de contours parasites qui peuvent, eux-mêmes, rendre l'extraction des contours de la bouche difficile. La thermographie a été suggérée comme une source d'information alternative pour la détection d'indices visuels sur des visages ou la reconnaissance d'individus.

La section 4.2 sera consacrée aux principes du rayonnement thermique. Nous rappellerons brièvement la définition du rayonnement thermique et son intérêt dans le cas de l'analyse labiale. Nous rappellerons les propriétés en émission du corps noir, ainsi que celles de la peau qui nous intéressent pour le cas de l'analyse labiale.

Dans la section 4.3, nous reviendrons plus en détail sur la bande infrarouge du spectre électromagnétique et sur les capteurs infrarouges dans le but de choisir un système d'acquisition pertinent pour l'analyse labiale.

Pour cette étude, l'absence de base de données proposant des images de visage, à la fois dans la modalité infrarouge et dans la modalité visible, nous a conduit à créer une base d'images combinée visible-infrarouge. La section 4.4 présentera la base d'images combinée

visible/infrarouge qui a été constituée pour cette étude, ainsi que le matériel employé pour en faire l'acquisition.

Enfin, dans la section 4.5, nous développerons une étude statistique du contraste peau-lèvre d'une manière analogue à celle qui a été conduite aux chapitres 1 et 2 sur des images de bouche en couleurs.

4.2 Rappels sur le rayonnement thermique

Tout objet dont la température est supérieure au zéro absolu émet des ondes électromagnétiques appelées rayonnement thermique ou plus familièrement « chaleur ». Cette émission d'ondes électromagnétiques est le résultat de l'agitation et des chocs entre les particules constituant l'objet. Cette agitation est mesurée par la température de l'objet. Le rayonnement thermique est caractérisé par des longueurs d'ondes comprises entre 0.1 et $100 \mu m$ ($1\mu m=10^{-6}m$). La figure 4.1 présente le spectre électromagnétique. On constate que les domaines, visible et infrarouge, occupent une petite portion du spectre.

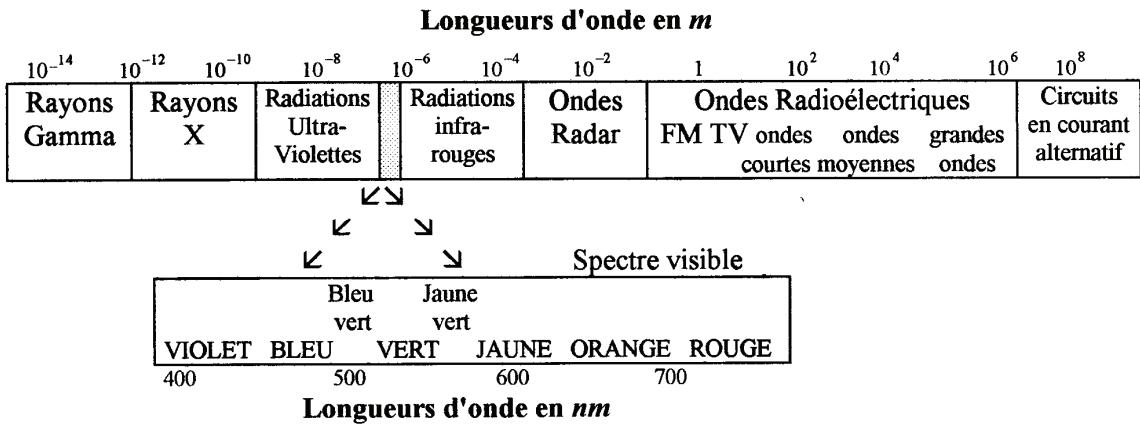


Figure 4.1 : Spectre électromagnétique

L'évaluation des propriétés en émission d'un corps réel quelconque se base sur les propriétés du corps noir. Le corps noir constitue le modèle de l'émetteur idéal qui rayonnerait un maximum d'énergie à chaque température et pour chaque longueur d'onde. On considère également le corps noir comme une surface lambertienne. C'est-à-dire qu'à une température et une longueur d'onde données, le corps noir rayonne de la même manière

dans toutes les directions. Un corps noir possède également la particularité d'être un absorbeur idéal. Le corps noir吸t toutes les ondes électromagnétiques qu'il reçoit sans en réfléchir ni en transmettre. La loi de Planck définit la distribution de luminance énergétique monochromatique du rayonnement thermique du corps noir en fonction de la température thermodynamique. La loi de Planck est donnée par l'expression suivante :

$$L_\lambda^\circ = \frac{2hc_\lambda^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc_\lambda}{\lambda kT}\right) - 1} W^2 m^{-3} sr^{-1} \quad (45)$$

où $c_\lambda=c/n_\lambda$ est la vitesse du rayonnement électromagnétique dans le milieu où se propage le rayonnement, n_λ est l'indice de réfraction du milieu pour la longueur d'onde λ ($n_\lambda \sim 1$ pour l'air), $c = 299\,792\,458\,ms^{-1}$ est la vitesse de la lumière dans le vide, $h=6.62617\times10^{-34}\,Js$ est la constante de Planck, $k=1.38066\times10^{-23}\,JK^{-1}$ est la constante de Boltzmann et T est la température de la surface du corps noir en Kelvin (K). À la figure 4.2, nous donnons le tracé de L_λ° pour différentes températures du corps noir. Nous avons tracé en particulier les courbes de rayonnement du corps noir pour les températures $T_{soleil} = 5777\,K$ et $T_{corps} = 310.15\,K$ qui correspondent respectivement aux températures à la surface du soleil et à la température du corps humain. Le corps noir est un modèle idéal, l'exemple le plus connu qui s'en rapproche est notre soleil. On constate sur la figure 4.2 que le spectre visible par le système visuel humain est centré autour de la longueur d'onde $\lambda \sim 0.5\,\mu m$ pour laquelle le soleil émet le maximum d'énergie.

En pratique, pour le cas d'un corps réel, les propriétés en émission sont obtenues par comparaison avec les propriétés du corps noir pour des conditions de température et de longueur d'onde identiques. Pour un objet quelconque, on définit alors un coefficient ε , appelé émissivité, qui correspond au rapport entre l'énergie rayonnée par l'objet et celle qui serait rayonnée par un corps noir dans les mêmes conditions. Pour un corps noir, l'émissivité est constante $\varepsilon_{cn}=1$. Pour un matériau quelconque, l'émissivité peut varier en fonction de la longueur d'onde, de la direction d'émission et de la température de surface.

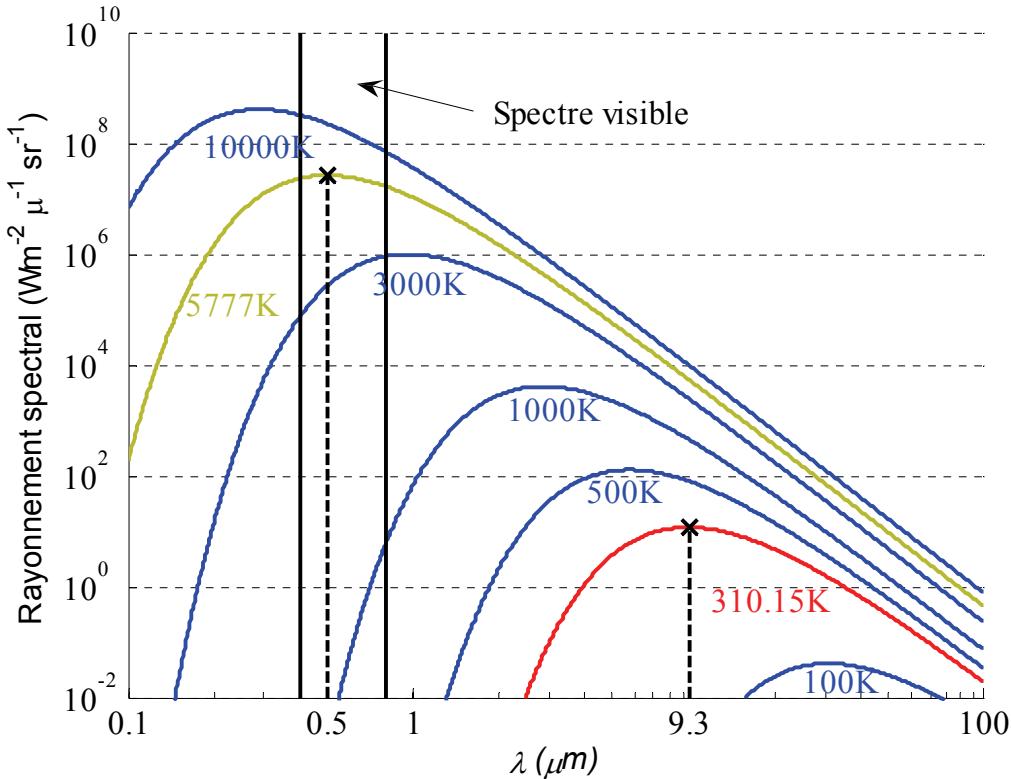


Figure 4.2 : Spectre de rayonnement de Planck pour différentes températures du corps noir. En rouge, nous avons tracé la courbe à $T=310.15\text{K}$ ce qui correspond à la température du corps humain. En jaune, nous avons tracé la courbe du rayonnement pour $T=5777\text{K}$ qui correspond approximativement à la température à la surface du soleil.

Dans (Steketee, 1973), l'auteur a effectué une étude de propriétés d'émission de la peau pour des sujets ayant différentes teintes de peau (blanche, noire, ...). L'auteur a constaté que la peau pouvait être considérée comme un corps gris diffusant (ϵ est indépendant de la longueur d'onde et de la direction) dont l'émissivité est égale à $\epsilon_{\text{peau}} = 0.98+-0.01$. On peut donc assimiler la peau à un corps noir. Sur la figure 4.2, nous avons tracé la courbe de l'énergie rayonnée par le corps noir à $T_{\text{corps}}=310.5\text{ K}$, ce qui correspond à la température du corps humain. On constate bien que, à $T=T_{\text{corps}}$, la peau n'émet pas dans le domaine visible du spectre. Le maximum d'énergie rayonnée est obtenu à la longueur d'onde $\lambda \sim 9.3\text{ }\mu\text{m}$. En se plaçant dans cette gamme de longueurs d'ondes pour observer le visage, il est possible d'observer le visage indépendamment des conditions d'éclairage. En effet, dans cette bande, on observe le rayonnement émis par la peau et non pas le rayonnement réfléchi par la peau, comme dans le domaine visible du spectre électromagnétique. A priori, en

travaillant autour de $\lambda \sim 9.3 \mu m$, les images de visage seront indépendantes des conditions d'illumination.

Dans cette thèse, nous sommes intéressés à segmenter les contours des lèvres sur des images de visage. L'objectif est de travailler avec des images qui offrent le maximum de contraste entre la peau et les lèvres. De plus, il faut avoir à l'esprit que l'air n'est pas transparent pour toutes les longueurs d'ondes comprises dans le domaine infrarouge. Dans la section suivante, nous nous proposons donc d'étudier plus en détail le domaine infrarouge et de déterminer la bande de fréquences et le système d'acquisition optimal pour l'analyse labiale.

4.3 Analyse du domaine infrarouge et choix d'un système d'acquisition adapté à l'analyse labiale

4.3.1 Spectre infrarouge

Dans la section précédente, nous avons évoqué les propriétés en émission de la peau. Nous avons pu voir que le corps, par l'intermédiaire de la peau, n'émettait pas de radiation thermique dans le domaine visible du spectre mais que l'essentiel de l'énergie rayonnée par le corps l'était dans la bande infrarouge qui s'étend entre 1 et $100 \mu m$. De plus, il faut avoir à l'esprit que les ondes électromagnétiques vont se propager dans l'air. L'air est un mélange complexe de gaz différents, de particules et autres aérosols avec des concentrations variables. Les différents composants de l'air ne sont pas transparents pour toutes les longueurs d'onde dans la bande infrarouge. L'air est principalement composé de gaz diatomiques (O_2 et N_2) qui sont transparents pour toutes les longueurs d'onde. Mais, on retrouve également des molécules triatomiques (O_3 , CO_2 , H_2O) qui ne sont pas transparentes pour toutes les longueurs d'onde :

- L'ozone (O_3) absorbe les longueurs d'onde pour $\lambda < 0.3 \mu m$.
- Le CO_2 absorbe les ondes électromagnétiques pour les longueurs d'ondes λ suivantes : $2.36 \mu m < \lambda < 3.02 \mu m$, $4.01 \mu m < \lambda < 4.80 \mu m$ et $12.5 \mu m < \lambda < 16.5 \mu m$
- L'eau (H_2O) absorbe les ondes de longueur d'onde λ : $2.24 \mu m < \lambda < 3.25 \mu m$, $4.80 \mu m < \lambda < 8.5 \mu m$ et $12 \mu m < \lambda < 25 \mu m$.

Le coefficient d'absorption de ces gaz est fonction de la pression partielle du gaz et de la distance traversée. La figure 4.3 présente la transmittance atmosphérique pour une atmosphère standard et pour une distance de 1 km. Dans le cas d'application en intérieur, avec une distance entre le sujet et la caméra inférieure à 2 mètres, l'absorption de l'atmosphère est néanmoins négligeable (Maldague, 2001).

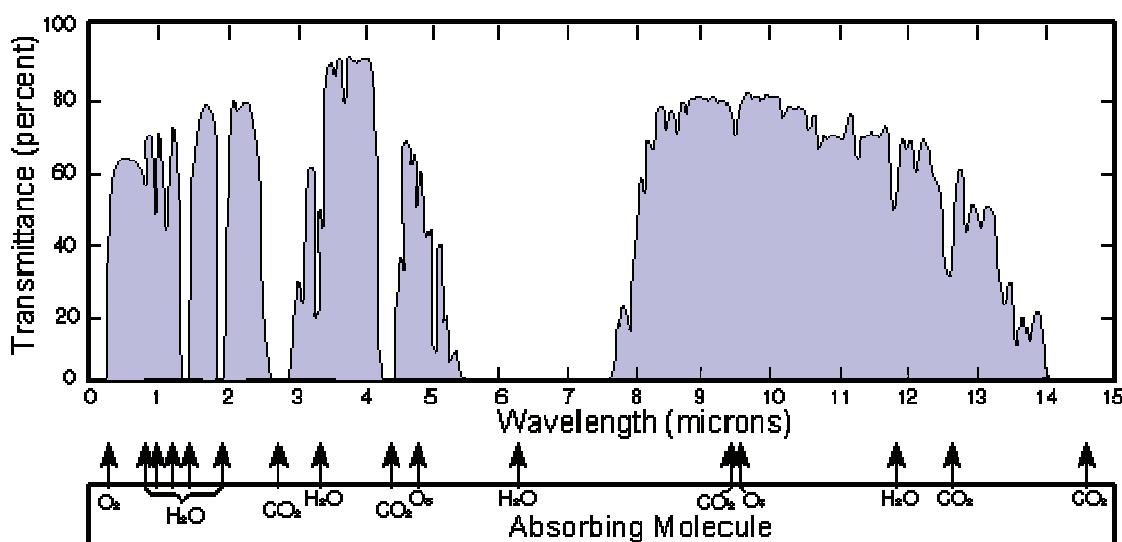


Figure 4.3 : Transmittance atmosphérique pour le domaine infrarouge avec une atmosphère standard aux États-Unis en 1976, au niveau de la mer, $T=288\text{K}$, taux d'humidité 46%, pression atmosphérique 1013 millibar, et sur un parcours d'un kilomètre horizontal. En abscisse, on donne la longueur d'onde en μm . En ordonnée, on donne la transmittance en %. Les gaz causant l'absorption sont également indiqués (Wikipedia, 2009).

4.3.2 Capteurs en imagerie infrarouge

En imagerie thermique, on distingue 2 grandes catégories de capteurs : les capteurs quantiques et les capteurs thermiques. Ces capteurs possèdent des bandes passantes différentes suivant la technologie employée, ce qui aura une influence sur les phénomènes que l'on souhaite observer. Dans le cas des capteurs thermiques, appelés aussi microbolomètres, le rayonnement thermique incident échauffe directement une matrice de pixels. Les pixels sont composés d'un matériau dont la résistance électrique se modifie en s'échauffant. La résistance est ensuite mesurée et convertie en température. La mesure de température est, a priori, indépendante de la longueur d'onde. La bande passante sera en

pratique limitée par les matériaux utilisés pour les optiques et par l'absorption des ondes dans l'atmosphère.

Les capteurs quantiques convertissent directement l'énergie reçue en signaux électriques. La réponse est dépendante de la fréquence de l'onde électromagnétique incidente. Les capteurs quantiques offrent des temps de réponse plus rapides que pour les capteurs thermiques ainsi que de meilleures performances en détection. Parmi les détecteurs quantiques, il existe 2 sous-catégories : les capteurs photovoltaïques et les capteurs photoconducteurs. Pour les capteurs photoconducteurs, le rayonnement électromagnétique augmente la conductivité du matériau. Pour les capteurs photovoltaïques le rayonnement entraîne la production d'électrons qui créent un courant électrique. Les capteurs quantiques sont, la plupart du temps, refroidis pour limiter leur propre bruit thermique. La table 4.1 résume les différents types de capteurs existants ainsi que leurs principales caractéristiques.

Type de capteur		Matériaux	Bande passante (μm)	Température de fonctionnement (K)
Thermique	Micro-bolomètres	a-Si VO _x	Dépend de l'optique, typiquement : 8-14	300
Quantique	Photoconducteurs	Pbs PbSe InSb HgCdTe	1 – 3.6 1.5 – 5.8 2 – 6 2 – 16	300 300 213 77
		Ge InGaAs Ex. InGaAs InAs InSb HgGdTe	0.8 – 1.8 0.7 – 1.7 1.2 – 2.55 1 – 3.1 1 – 5.5 2 – 16	300 300 253 77 77 77

Table 4.1 : Capteurs les plus utilisés en imagerie infrarouge.

La figure 4.4 représente les tracés de la détectivité en fonction de la longueur d'onde λ pour les capteurs de la table 4.1. La détectivité D^* correspond à la sensibilité du capteur par unité de surface et pour une largeur de bande normalisée de 1Hz.

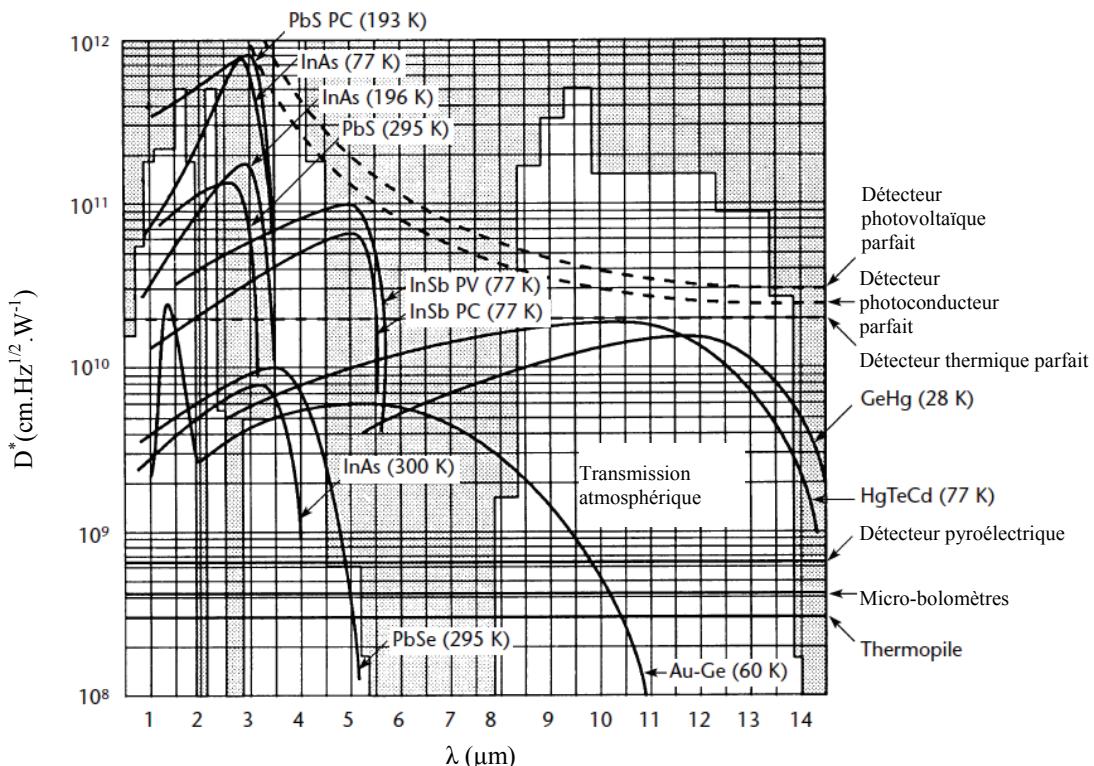


Figure 4.4 : Déetectivité D^* en fonction de la longueur d'onde λ pour différents types de capteurs infrarouges (Maldaige, 2001).

La figure 4.4 montre que, globalement, D^* est plus important pour les détecteurs quantiques que pour les détecteurs thermiques. Pour les détecteurs thermiques, la sensibilité est beaucoup moins importante que pour les détecteurs quantiques. On peut également rajouter que, d'une manière générale et même si cela va dépendre du système d'acquisition complet, les détecteurs quantiques ont des temps de réponse inférieurs à ceux des micro-bolomètres. Pour les détecteurs quantiques, on peut typiquement obtenir des temps de réponse de l'ordre de $1\mu\text{s}$. Pour les microbolomètres, l'ordre de grandeur est proche de 10ms . En ce qui concerne le prix, le coût de fabrication des microbolomètres est moins élevé que pour les détecteurs quantiques. Le fait que ces capteurs fonctionnent à température ambiante permet de fabriquer des systèmes d'imagerie moins chers et accessibles au grand public. Pour ce qui est des capteurs quantiques, pour obtenir la meilleure sensibilité, le refroidissement est nécessaire (figure 4.4), ce qui implique une augmentation sensible du coût des systèmes.

d'imagerie utilisant ces capteurs. On retrouvera plus ce type de capteurs dans des caméras de recherche et développement.

Compte tenu des contraintes imposées par les bandes d'absorption de l'atmosphère et les bandes passantes des capteurs, le domaine infrarouge a été subdivisé de la manière suivante :

- Proche infrarouge (Near Infrared, *NIR*) : De $0.7 \mu\text{m}$ à $1.0 \mu\text{m}$. Cette bande s'étend approximativement depuis la fin de la bande passante de l'œil jusqu'à à la fin de celle du silicium.
- Infrarouge court (Short-wave Infrared, *SWIR*): De $1 \mu\text{m}$ à $3 \mu\text{m}$. Cette bande s'étend de la fin de la bande passante du silicium jusqu'à la première bande d'absorption du CO₂.
- Infrarouge moyen (Mid-wave Infrared, *MWIR*) : De $3 \mu\text{m}$ à $5 \mu\text{m}$. Ce domaine correspond à une fenêtre atmosphérique (voir figure 4.3) et est couvert par les détecteurs InSb, HgCdTe et PbSe.
- Infrarouge lointain (Long-wave Infrared, *LWIR*) : De $8 \mu\text{m}$ à $14 \mu\text{m}$. Fenêtre atmosphérique couverte par les détecteurs HgCdTe et les microbolomètres.
- Infrarouge très-lointain (Very-long Wave Infrared, *VLWIR*) : De $14 \mu\text{m}$ à $30 \mu\text{m}$. Application spécifique incluant des lasers avec de grandes longueurs d'onde. Bande couverte par des détecteurs spécifiques au silicium dopé.

4.3.3 Choix d'un système d'acquisition infrarouge pour l'analyse labiale

Dans la section 4.2, nous avons étudié les propriétés en émission de la peau et le rayonnement du corps noir pour une température $T_{corps} = 310.15\text{K}$. Sur la figure 4.2, nous avons constaté que le spectre d'énergie rayonnée s'étale autour de $\lambda=9.3 \mu\text{m}$. Pour effectuer nos observations du visage, il faudra donc chercher à travailler dans des bandes de longueurs d'onde voisines de $\lambda = 9.3\mu\text{m}$. Dans notre cas, nous avons eu à disposition des caméras avec des capteurs quantiques refroidis opérant dans la bande $3-5\mu\text{m}$ et des caméras à microbolomètres opérant dans la bande $8-14\mu\text{m}$. Pour la bande $3-5\mu\text{m}$, la caméra dispose d'un capteur InSb refroidi à 77k, de résolution 512x640, et elle fournit des images avec une dynamique codée sur 14 bits. Pour la bande $8-14\mu\text{m}$, nous disposions d'une caméra à

microbolomètres non refroidi avec un capteur de résolution 160x128. D'après les éléments présentés aux sections 4.3.1 et 4.3.2, la bande infrarouge moyen (*MWIR*, 3-5 μm) et la bande infrarouge lointain (*LWIR*, 8-14 μm) sont toutes les 2 exploitables pour analyser le visage. On se propose d'étudier l'intérêt de chacune de ces 2 bandes de fréquences du point de vue de l'analyse du visage. Si l'on considère la peau comme un corps gris diffusant dont l'émissivité est égale à $\varepsilon_{peau} \sim 1$, nous pouvons alors calculer la luminance énergétique émise dans les 2 bandes infrarouges en intégrant la loi de Planck :

$$R_{3-5\mu m}^{\circ} = \int_{3 \cdot 10^{-6}}^{5 \cdot 10^{-6}} \frac{2hc_{\lambda}^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc_{\lambda}}{\lambda kT}\right) - 1} d\lambda = 2.65 \text{ } W^2 m^{-2} sr^{-1} \quad (46)$$

$$R_{8-14\mu m}^{\circ} = \int_{8 \cdot 10^{-6}}^{14 \cdot 10^{-6}} \frac{2hc_{\lambda}^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc_{\lambda}}{\lambda kT}\right) - 1} d\lambda = 63.83 \text{ } W^2 m^{-2} sr^{-1} \quad (47)$$

Du point de vue de l'intensité du rayonnement, il sera donc plus intéressant de travailler dans la bande 8–14 μm . Cette bande de fréquence sera particulièrement appropriée pour des applications de détection impliquant des sujets humains dans des environnements extérieurs ou difficiles (atmosphère humide, présence de particules en suspension). Pour l'analyse labiale, nous chercherons à obtenir des images proposant le plus grand contraste entre les zones de peau et les lèvres sur des images de visage acquises à de faibles distances. Dans le cas d'images thermiques, le système d'acquisition mesure le rayonnement émis par la peau, rayonnement qui est fonction de la température. Le contraste thermique sur le visage sera donc un paramètre important pour l'analyse labiale. Le contraste thermique $C_{\Delta\lambda}$ s'exprime de la manière suivante :

$$C_{\Delta\lambda} = \frac{R_{\Delta\lambda}^{\circ}(T_2) - R_{\Delta\lambda}^{\circ}(T_1)}{R_{\Delta\lambda}^{\circ}(T_2) + R_{\Delta\lambda}^{\circ}(T_1)} \quad (48)$$

$$R_{\Delta\lambda}^{\circ} = \int_{\Delta\lambda} \frac{2hc_{\lambda}^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc_{\lambda}}{\lambda kT}\right) - 1} d\lambda \text{ } W^2 m^{-2} sr^{-1}$$

La figure 4.5 présente les tracés des courbes du contraste thermique $T_2 - T_1 = 1\text{K}$ pour différentes températures T_2 ainsi que le contraste pour différents $\Delta(\text{K}) = T - T_{\text{corps}}$ autour de $T_{\text{corps}}=310.15\text{K}$.

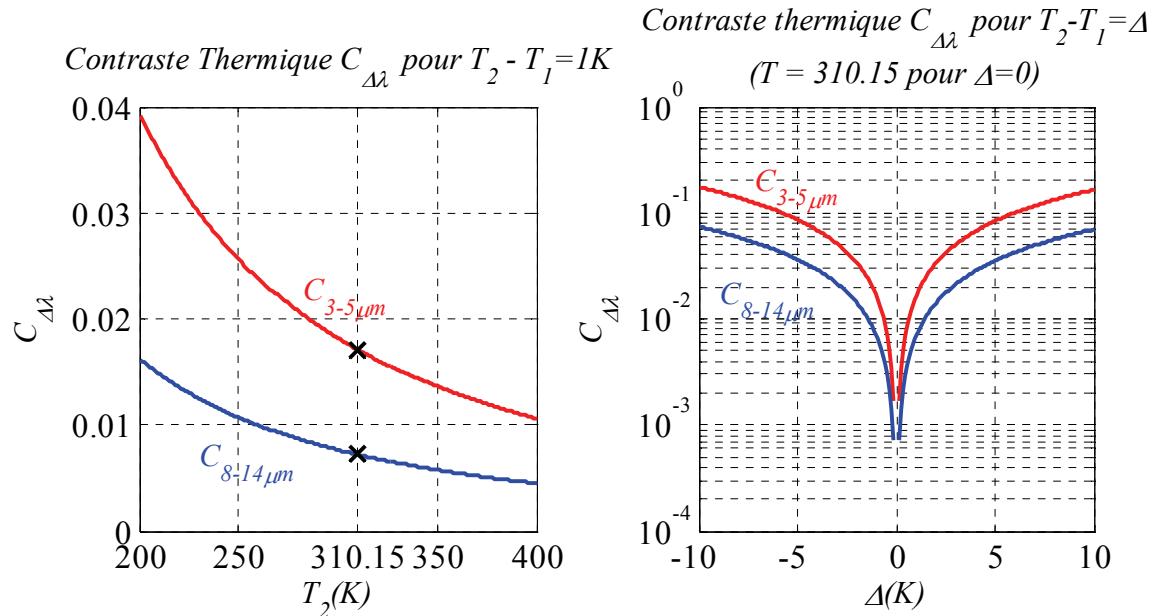


Figure 4.5 : Contraste thermique dans les bandes *MWIR* et *LWIR*, à gauche, nous avons tracé le contraste pour $T_2 - T_1 = 1\text{K}$, à droite, nous avons tracé le contraste pour des $\Delta(\text{K})$ autour de $T_{\text{corps}}=310.15\text{K}$.

Les tracés de la figure 4.5 montrent que le contraste thermique est meilleur dans la bande *MWIR* ($3-5\mu\text{m}$). Sur la tracé de gauche, pour $T = 310.15\text{K}$, le ratio $C_{3-5\mu\text{m}}/C_{8-14\mu\text{m}} = 2.38$. Le tracé de droite donne l'évolution du contraste thermique pour des Δ (en K) autour de 310.15K . On constate également que le contraste thermique est plus important dans la bande *MWIR*. Il faut également se rappeler que la sensibilité du capteur InSb, dans la bande $3-5\mu\text{m}$, est bien plus importante que pour les microbolomètres dans la bande $8-14\mu\text{m}$. La déetectivité est de l'ordre de $4.10^{10}\text{cm.Hz}^{1/2}.W^{-1}$ pour le capteur InSb dans la bande $3-5\mu\text{m}$ alors que pour les microbolomètres la déetectivité est de l'ordre de $4.10^8\text{cm.Hz}^{1/2}.W^{-1}$. La caméra avec le capteur InSb refroidi est donc bien plus sensible que la caméra à microbolomètre. De plus, il faut souligner que la caméra à microbolomètre n'étant pas refroidie, la caméra doit être calibrée périodiquement pour compenser le réchauffement propre du capteur. Pour le problème de la séparation peau/lèvre, l'acquisition des images se faisant à des distances relativement faibles (inférieures à 5 m) et en intérieur, les

perturbations atmosphériques sont faibles. Au vu des spécifications des caméras dont nous disposons et des caractéristiques du rayonnement thermique du visage, il nous a semblé pertinent d'utiliser la caméra possédant le capteur InSb. Dans la bande $3\text{-}5\mu\text{m}$, l'énergie rayonnée est certes plus faible que dans la bande $8\text{-}14\mu\text{m}$ mais le contraste thermique est plus fort et la caméra avec le capteur InSb refroidi à 77K est bien plus sensible que la caméra à microbolomètres. La résolution du capteur InSb est également bien plus importante que le capteur de la caméra à microbolomètres, ce qui est également un paramètre important dans notre cas où la précision est un objectif.

En conclusion, il y aura des compromis à faire pour choisir le système d'acquisition, et donc la bande de fréquence pertinente, pour étudier une source de rayonnement thermique. Dans notre cas, nous étudions la séparation peau/lèvre, donc nous cherchons à avoir le plus grand contraste sur le visage et la plus grande précision. Nous avons donc privilégié la bande *MWIR* pour étudier la séparation peau/lèvre.

4.4 Crédation d'une base d'image visible/infrarouge

Dans les sections précédentes de ce chapitre, nous avons étudié les propriétés en émission de la peau et les contraintes particulières liées au travail dans la bande infrarouge pour l'analyse labiale (bandes adaptées à l'étude des lèvres, réponses des capteurs, perturbations liées à l'atmosphère). La plupart des travaux d'analyse faciale tels que ceux présentés dans (Yoshitomi, 1997; Socolinsky, 2003) ont choisi de faire l'acquisition des images avec des systèmes d'acquisition travaillant dans la bande *LWIR* ($8\text{-}14\mu\text{m}$). Nous avons, quant à nous, privilégié la bande *MWIR* (cf. section 4.3.3). L'absence de base de données, proposant à la fois des images de visage dans la modalité infrarouge et dans la modalité visible, nous a conduit à créer une base d'images combinée visible/infrarouge.

Lors de la création de cette base, l'objectif visé était d'offrir un ensemble d'images de visage, dans la modalité visible et la modalité infrarouge (bande *MWIR*, $3\text{-}5\mu\text{m}$), destiné à servir de base de test pour des applications d'analyse faciale. Nous voulions pouvoir comparer les images visibles et infrarouges et permettre la fusion entre les modalités. Les contraintes auxquelles nous avons été confrontés pour l'acquisition des images ont été les suivantes :

- Acquisition simultanée et synchrone des séquences dans les modalités visible et infrarouge.
- Les vues des 2 caméras devaient être les plus proches possibles pour autoriser la fusion des images visibles et infrarouges.

Notre base est constituée de séquences d'images de 17 sujets différents (15 hommes et 2 femmes). L'ensemble des sujets inclut aussi bien des individus sans signes distinctifs que des individus présentant des indices visuels comme des lunettes, une barbe, une moustache, Les sujets ont été filmés dans un environnement fermé avec un éclairage de type fluorescent sur un fond blanc mat. Les tubes fluorescents présentent l'avantage de ne pas émettre dans la bande infrarouge. Avec ce type d'éclairage, il n'y a pas de perturbations au moment de l'acquisition avec une caméra infrarouge. Au contraire, un éclairage par ampoule à incandescence peut engendrer des perturbations, notamment des réflexions parasites (ex. réflexions sur les lunettes). Pour chaque sujet, 2 séquences de 400 images ont été filmées de manière synchrone. Une séquence était filmée avec une caméra visible et une séquence était filmée avec une caméra infrarouge. On dispose au total de 6800 images dans chaque modalité. Les images ont été cadrées sur le visage de manière à ce que celui-ci soit bien centré et que les cheveux soient visibles. Au cours des acquisitions, il a été demandé à chaque sujet de compter de 1 à 10 en français. La figure 4.6 présente des images tirées de la base pour les 2 modalités.

Pour faire l'acquisition simultanée des images dans les 2 modalités, nous avons utilisé un système de paire stéréo avec un signal de synchronisation provenant d'une source externe. Les 2 caméras étaient montées sur un bras côté à côté de manière à ce que les positions des capteurs soient identiques. Nous avons employé des objectifs avec une distance focale fixe de 50mm dans les 2 cas. Les 2 caméras ont été configurées de manière à déclencher la capture d'une image à partir d'un signal TTL avec une fréquence de 30 Hz. Pour faire l'acquisition dans la modalité infrarouge nous avons utilisé la caméra infrarouge de R&D FLIR Systems ThermaCAM TM Phoenix. Cette caméra dispose d'un capteur de type InSb refroidi opérant dans la bande $3\text{-}5\mu\text{m}$. La résolution du capteur est de 512x640 pixels. Les images fournies voient leurs dynamiques codées sur 14 bits. Pour pouvoir afficher les images de la figure 4.6, la dynamique de chaque image a été centrée sur la distribution du

visage. Une conversion a été effectuée pour obtenir des images dont le niveau de gris est codé sur 8 bits. Le temps d'intégration de la caméra infrarouge a été fixé de manière à ce qu'il n'y ait pas de saturation sur la zone du visage.

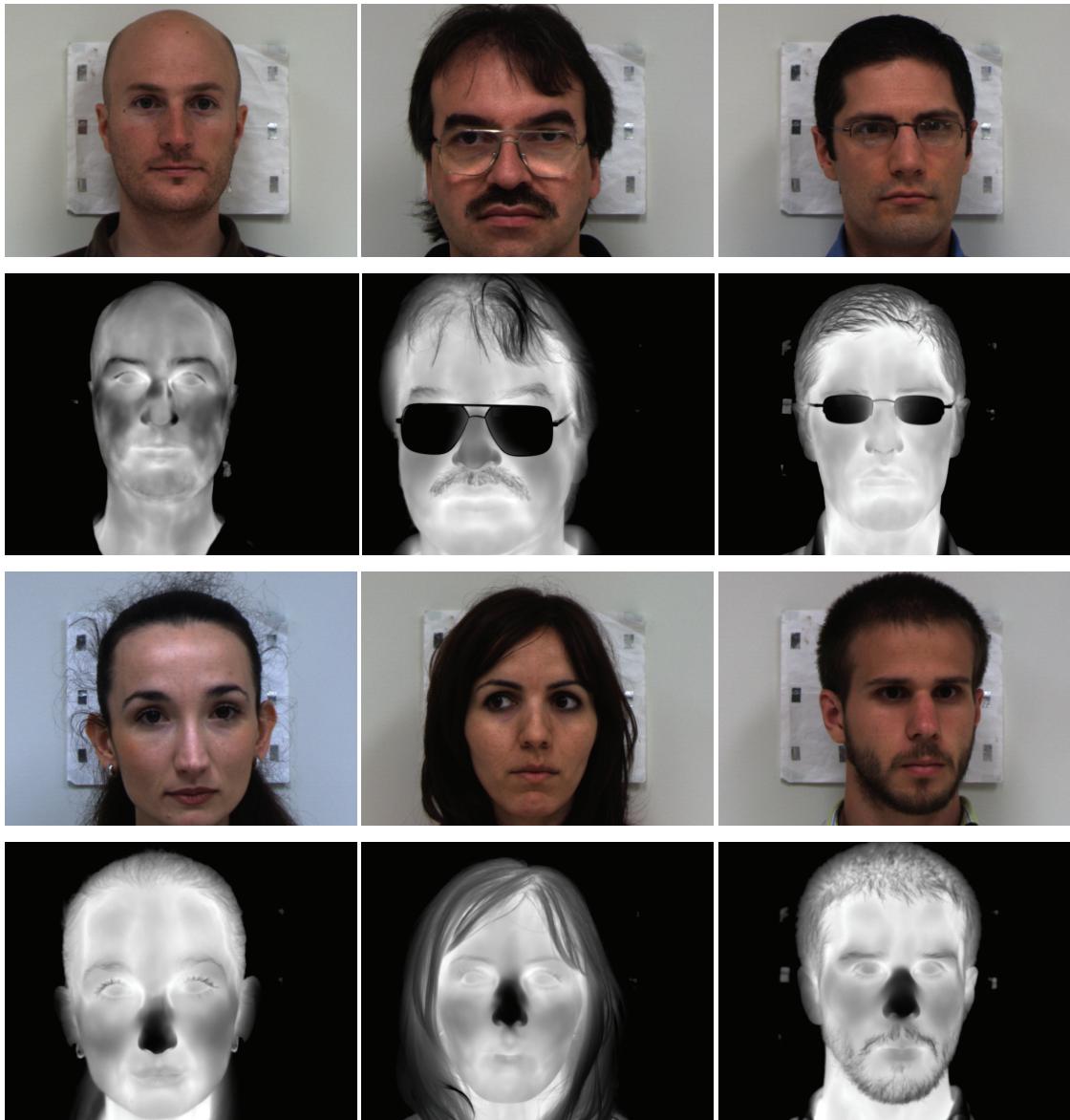


Figure 4.6 : Exemples d'images tirés de la base conjointe visible/infrarouge

Pour l'acquisition dans la modalité visible, une caméra mono-CCD de vision, avec une résolution de capteur de 1032x776, a été employée. Cette caméra autorise la synchronisation de la capture sur un signal externe. Pour obtenir des images en couleurs

avec un signal de synchronisation à 30Hz, la résolution des images capturées a été fixée à 816x590 pixels. Le temps d'intégration du capteur a été ajusté de manière à ce qu'il n'y ait pas de pixel saturé sur le visage.

La figure 4.7 présente un schéma du système assemblé pour faire l'acquisition de la base d'image.

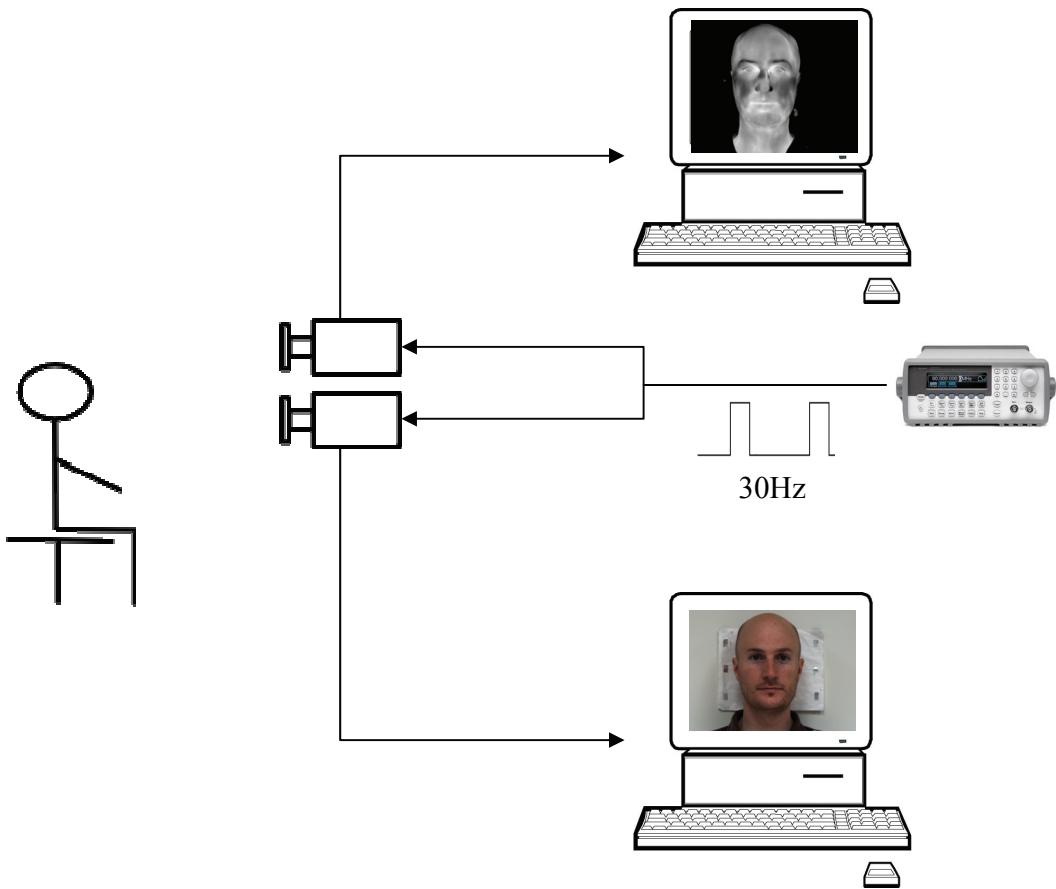


Figure 4.7 : Schéma du système d'acquisition de la base d'image de visage visible/infrarouge

Chaque caméra a été reliée à un ordinateur particulier pour l'enregistrement des séquences. Pour synchroniser les captures sur les 2 caméras, nous avons utilisé un générateur de fonction programmé pour générer un signal TTL à 30 Hz avec 400 impulsions. Les caméras étaient programmées pour capturer une image après détection d'un front montant sur le signal de synchronisation externe. Les acquisitions ont été faites sans calibration préalable de la paire stéréo. La calibration aurait nécessité la pose ou la projection sur le visage des

sujets d'indices visuels visibles dans les 2 bandes (visible et infrarouge) ce qui était difficile en pratique. Notre objectif étant d'étudier la zone de la bouche une localisation manuelle a été privilégiée pour extraire la zone de la bouche dans les 2 modalités (cf. section 4.5.1).

4.5 Etude du contraste peau/lèvre dans la modalité infrarouge

4.5.1 Mise en registre des images de bouche

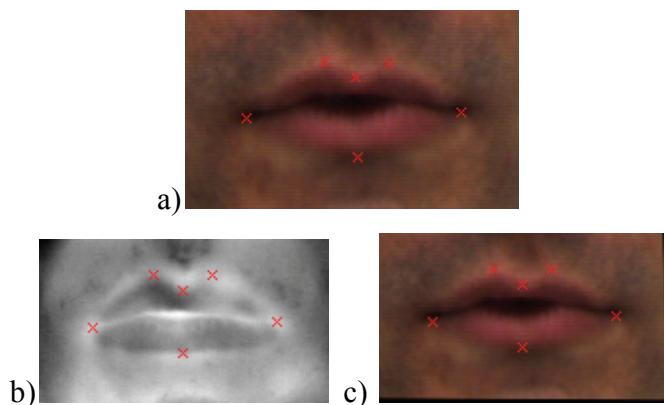


Figure 4.8 : Annotation d'une image de bouche dans la modalité visible et la modalité infrarouge avec 6 paires de points pour la mise en registre, a) Annotation de l'image visible de départ, b) Annotation de l'image infrarouge, c) Image visible transformée.

Nous disposons de séquences dans le domaine visible et dans le domaine infrarouge de 400 images chacune pour 17 sujets. Les séquences ayant été acquises en environnement contrôlé, nous avons choisi d'annoter manuellement les contours internes et externes de la bouche sur 5 images pour chaque modalité et pour chaque sujet afin de mener à bien notre étude. Sur ces 5 images, nous avons sélectionné manuellement 2 images où la bouche est fermée, une au début de la séquence et une en fin de séquence. Pour les 3 autres images, nous avons sélectionné des images où la bouche est ouverte au hasard dans les séquences du sujet. Dans la section précédente, nous avons précisé les conditions dans lesquelles ont été faites les acquisitions dans les 2 modalités. En particulier, si les images ont été capturées simultanément dans les 2 modalités, la résolution des images est différente dans les 2 modalités ainsi que les angles de prise de vue. La première étape de notre étude sur le contraste peau/lèvre a été d'effectuer une mise en registre des images sélectionnées. Tout d'abord, la région de la bouche a été segmentée et nous avons ensuite utilisé 6 paires de

points pour calculer la transformation affine entre les 2 images (figure 4.8-a et 4.8-b). Les images acquises dans la bande infrarouge ayant la résolution la plus faible, nous avons calculé la transformation de l'image visible vers l'image infrarouge. Le choix des points pour calculer la transformation est intuitif. Nous avons utilisé les 2 commissures, qui sont facilement identifiables dans les 2 modalités, 3 points sur l'arc de cupidon et le point se trouvant sur le contour externe inférieur ayant la même abscisse que le point central de l'arc de cupidon (figure 4.8). La figure 4.8-c présente une image visible transformée dans la base de l'image infrarouge. On distingue les effets de bord sur l'image. Ils correspondent aux pixels qui n'ont pu être évalués lors de la transformation. Cette opération de mise en registre a été répétée pour toutes les images sélectionnées pour l'étude du contraste peau/lèvre.

4.5.2 Segmentation manuelle des contours internes et externes des lèvres

Après avoir effectué la mise en registre des images, nous avons procédé à une nouvelle annotation manuelle des images pour segmenter les lèvres.

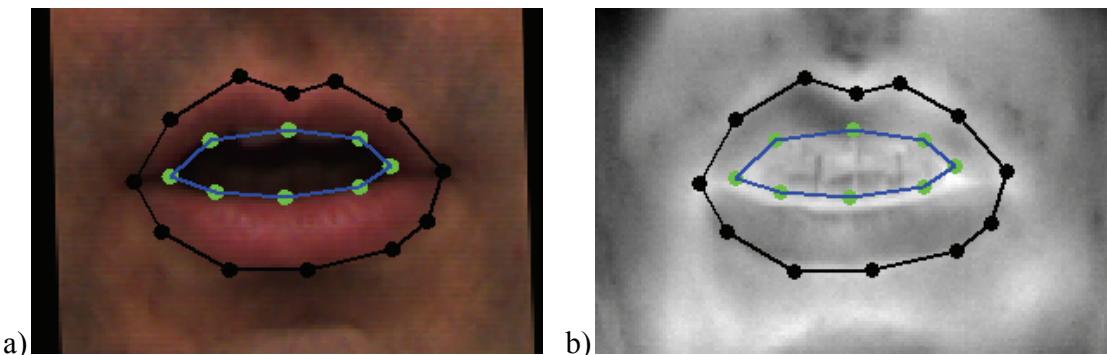


Figure 4.9 : Segmentation manuelle des lèvres pour le cas d'une bouche ouverte. En noir, nous avons tracé le contour externe et en bleu le contour interne.

A partir des images visibles transformées dans le repère des images infrarouges, nous avons segmenté le contour des lèvres en distinguant les cas « bouche ouverte » et « bouche fermée ». Pour le cas d'une bouche fermée, le contour externe seul a été segmenté en utilisant 12 points : Pour le contour externe supérieur, nous avons annoté les 2 commissures, 3 points pour l'arc de cupidon et 2 points intermédiaires pour caractériser la

courbure du contour. Lorsque la bouche est ouverte le contour externe est annoté de la même manière que dans le cas de la bouche fermée. Pour le contour interne, 8 points de contrôle ont été employés : 2 points correspondant aux « commissures internes », 3 points pour caractériser le contour interne supérieur et 3 points également pour caractériser le contour interne inférieur. Nous avons tracé sur la figure 4.9 les contours internes et externes segmentés manuellement pour le cas d'une bouche ouverte extraite de notre base.

4.5.3 Contraste peau/lèvre dans la modalité infrarouge

A partir des images annotées manuellement nous avons effectué une étude similaire à celle des sections 1.2 et 2.3. Nous avons extrait les ensembles de pixels des lèvres et de la peau pour la modalité infrarouge et pour la modalité visible. La table 4.2 donne les variances intraclasses et interclasses ainsi que le rapport V_{intra}/V_{inter} pour la modalité infrarouge (*IR*) et pour les différentes composantes couleur après traitement des images de la modalité visible par l'algorithme allongement-décorrélation. On constate que pour les grandeurs colorimétriques les résultats sont similaires à ceux de la section 2.3.

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
<i>IR</i>	$1.36 \cdot 10^{-3}$	$3.28 \cdot 10^{-5}$	41.64
R_{decorr}	$1.19 \cdot 10^{-2}$	$4.38 \cdot 10^{-3}$	2.71
G_{decorr}	$2.8 \cdot 10^{-3}$	$3.24 \cdot 10^{-3}$	0.86
B_{decorr}	$2.67 \cdot 10^{-3}$	$1.34 \cdot 10^{-4}$	19.918
Cb_{decorr}	$7.2 \cdot 10^{-4}$	$1.81 \cdot 10^{-4}$	3.97
Cr_{decorr}	$3.31 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$	1.0562
H_{decorr}	$6.85 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	2.7373
\hat{H}_{decorr}	$7.9 \cdot 10^{-3}$	$1.26 \cdot 10^{-2}$	0.62
\hat{U}_{decorr}	$3.66 \cdot 10^{-2}$	$5.03 \cdot 10^{-2}$	0.72

Table 4.2 : Variance intraclasses, Variance interclasses et V_{intra}/V_{inter} pour la modalité infrarouge *IR* et les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , \hat{H}_{decorr} et \hat{U}_{decorr} .

Nous avons également tracé les histogrammes normalisés des ensembles de pixels de la peau et des lèvres segmentés manuellement pour la modalité infrarouge à la figure 4.10. Pour chaque image, la dynamique a été centrée sur le niveau moyen de la peau pour égaliser les niveaux de rayonnement thermique émis par les différents sujets. Par la suite, la

dynamique des ensembles de pixels de la peau et des lèvres a été normalisée entre 0 et 1 pour les tracés de la figure 4.10.

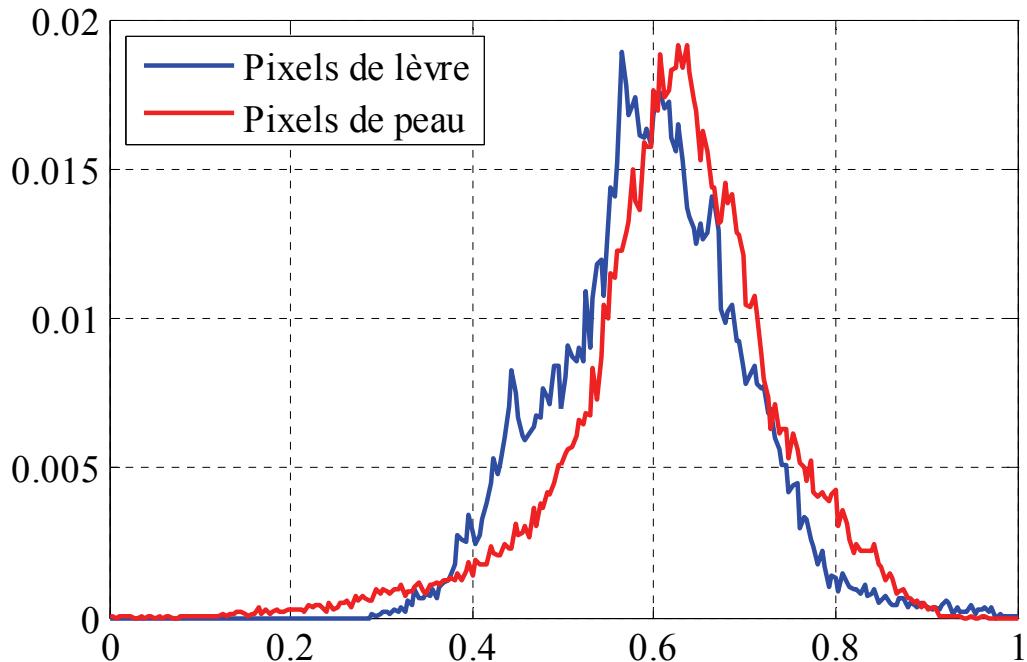


Figure 4.10 : Histogrammes normalisés des ensembles de pixels de la peau et des lèvres segmentés manuellement pour la modalité infrarouge.

Les résultats de la table 4.2 et les tracés des histogrammes de la figure 4.10 montrent un très fort recouvrement des distributions des ensembles de pixels de la peau et des lèvres. Les différences de rayonnement thermique sont très faibles entre la peau et les lèvres. Ces résultats indiquent que le contraste entre la peau et les lèvres est faible dans la modalité infrarouge. Il sera très difficile d'employer la modalité infrarouge pour localiser et modéliser le contour externe de la bouche. Pour illustrer les résultats de la table 4.2 et de la figure 4.10, nous présentons à la figure 4.11 des images de bouche dans la modalité infrarouge extraites de notre base ainsi que les images de teinte \hat{U} et l'image RGB correspondantes. Pour les sujets des 2 premières lignes, on constate qu'il y a très peu de contraste entre la peau et les lèvres par rapport à la composante \hat{U}_{decorr} . Pour le troisième sujet, le contraste peu/lèvre est meilleur. La lèvre inférieure, en particulier, possède une

température relativement homogène, inférieure à la peau, et le contour est assez net. Pour la lèvre supérieure, en revanche, le contour externe supérieur est presque invisible même si la température est légèrement inférieure au centre de la lèvre par rapport à la peau. Sur les 3 images infrarouges, on distingue un fort contraste entre les lèvres et la cavité buccale. Ce phénomène s'explique par le fait que ces images de bouches ouvertes ont été prises pendant que les sujets étaient en train d'expirer de l'air, air provenant des poumons. L'air expiré est à une température voisine du celle du corps et il faut se rappeler que, dans la bande $3-5\mu m$, il y a des bandes de fréquences pour lesquelles le CO₂ et les molécules d'eau ne sont pas complètement transparents (cf. figure 4.3). On distingue en fait le rayonnement émis par le CO₂ et l'eau (H₂O) contenus dans l'air expiré par les sujets.

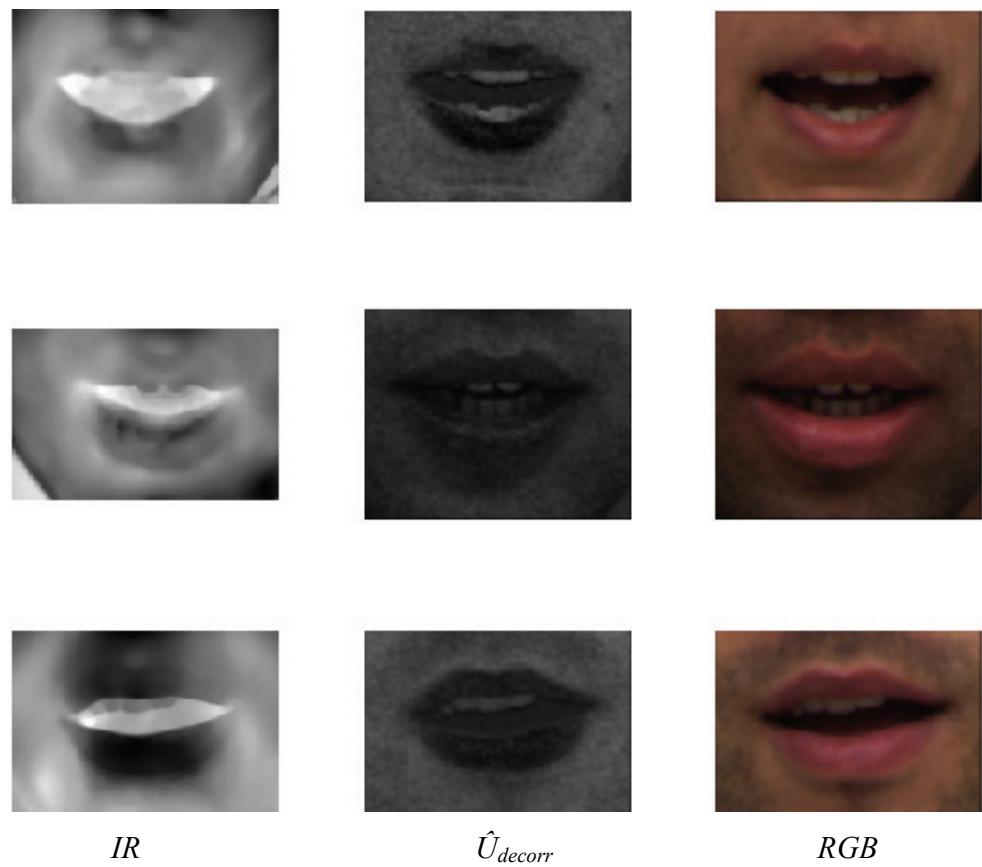


Figure 4.11 : Exemples d'images de bouches ouvertes dans la modalité infrarouge (IR) et dans la modalité visible (\hat{U} et RGB).

Dans la figure 4.12, nous présentons des images dans le cas où la bouche est fermée pour les mêmes sujets qu'à la figure 4.11. Sur les images infrarouges de la figure 4.12, on remarque que la température des lèvres reste stable par rapport au cas des bouches ouvertes.

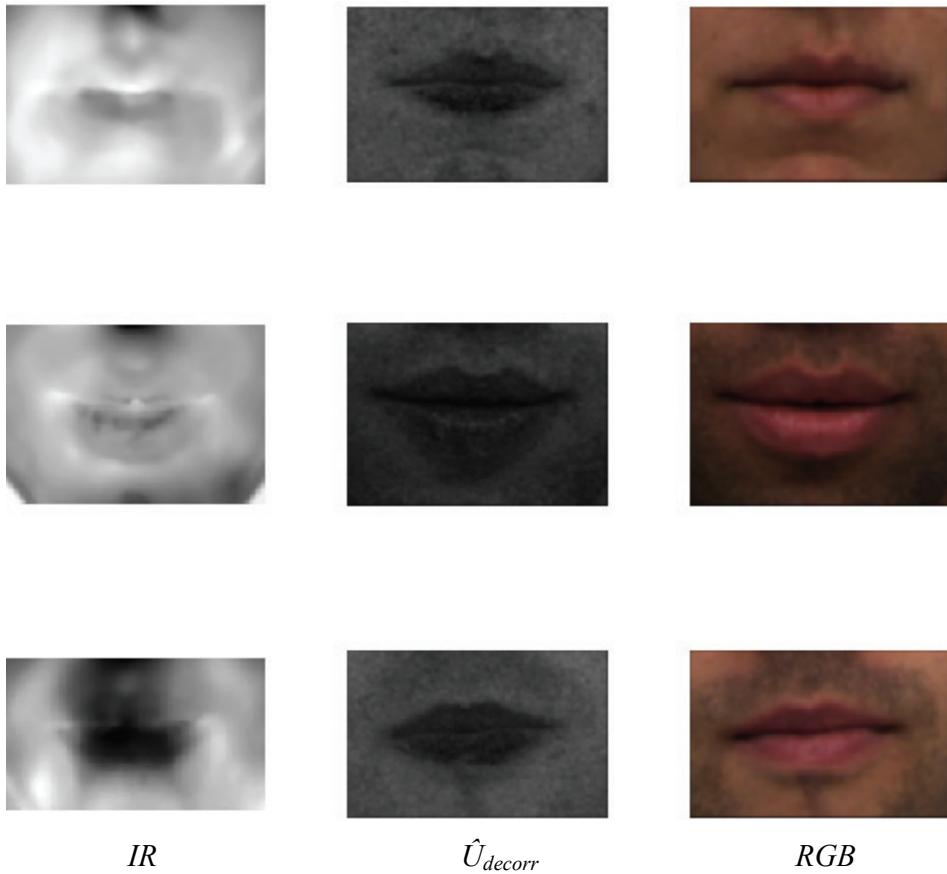


Figure 4.12 : Exemples d’images de bouches fermées dans la modalité infrarouge (IR) et dans la modalité visible (\hat{U} et RGB).

Les images de la figure 4.11 et 4.12 mettent en lumière la différence de nature entre l’information véhiculée par la modalité infrarouge et la modalité visible. L’image infrarouge nous apporte une information de type anatomique qui nous renseigne sur les caractéristiques du visage sous la peau (irrigation sanguine). Ce type d’information, très particulier à une personne, sera plutôt intéressant pour l’identification. On constate que, pour les 3 sujets présentés aux figures 4.11 et 4.12, les différences sont très importantes sur la zone de la bouche, mais que le contraste peau/lèvres est faible. L’information véhiculée par la modalité infrarouge possède un potentiel biométrique évident.

Pour la modélisation des lèvres l'information n'est, toutefois, pas assez stable d'un sujet à l'autre et surtout, la peau et les lèvres ne forment pas des ensembles suffisamment homogènes pour permettre de modéliser la bouche de manière robuste. L'information apportée dans le domaine visible concerne les propriétés de réflexion de la peau. Les propriétés des ensembles des pixels de la peau et des lèvres sont beaucoup plus stables, à éclairage constant, dans le domaine visible du spectre (cf. section 2.3, table 4.2) que les caractéristiques anatomiques visibles avec les images infrarouge.

4.6 Bilan

Dans ce chapitre, nous avons proposé une étude sur la modalité infrarouge pour l'analyse labiale. L'idée de départ était d'utiliser l'information apportée par la modalité infrarouge pour améliorer la robustesse de la détection des lèvres sur les visages. Pour cela, nous avons créé une base combinée d'images de visage dans les 2 modalités. L'objectif principal qui a guidé la création de la base était l'étude de la séparation peau/lèvres dans la modalité infrarouge. L'objectif secondaire était de proposer une base d'images permettant la fusion d'information, ce qui a imposé des contraintes de synchronisation sur les acquisitions. Nous avons précisé les critères de sélection du système d'acquisition pour la modalité infrarouge. Ces critères résultent d'un compromis entre les propriétés du rayonnement thermique du visage et les performances des caméras infrarouges disponibles. Dans notre cas, nous avons choisi une caméra infrarouge opérant dans la bande $3\text{-}5\mu\text{m}$. Si l'énergie rayonnée par le corps est moindre dans cette bande, le contraste thermique est plus fort et la caméra que nous avons utilisée offrait de meilleures performances que celle opérant dans la bande $8\text{-}14\mu\text{m}$ dont nous disposions. Finalement, l'étude de la séparation peau/lèvres dans la modalité infrarouge ne s'est pas révélée concluante. Par contre, nous avons pu constater le potentiel de cette modalité pour des problématiques de reconnaissance. Le rayonnement thermique émis par le sujet contient une information de type anatomique qui présente un fort potentiel biométrique.

Chapitre 5. Segmentation des contours externes et internes de la bouche

5.1 Introduction

Dans ce chapitre, nous abordons la partie finale de nos travaux sur la segmentation labiale, la segmentation des contours externes et internes des lèvres. Dans les chapitres précédents, nous avons présenté une méthode pour segmenter un masque binaire des lèvres (cf. chapitre 2) ainsi qu'une méthode pour déterminer l'état ouvert ou fermé de la bouche (cf. chapitre 3). Au chapitre 4, nous avons étudié l'intérêt de la modalité infrarouge pour séparer la peau et les lèvres. Cette étude ne s'est pas révélée concluante quant à la pertinence de la modalité infrarouge pour le problème de la séparation peau/lèvres, nous n'utiliserons donc pas la modalité infrarouge dans les traitements présentés dans ce chapitre. A partir de maintenant, nous supposons que l'on dispose du masque binaire des lèvres et aussi que l'on connaît l'état ouvert ou fermé de la bouche. Le masque binaire des lèvres nous fournit une description grossière de l'ensemble du contour externe des lèvres. Le masque binaire n'est, par contre, pas suffisamment précis pour nous renseigner sur la nature du contour interne des lèvres. Lorsque la bouche est fermée le contour interne se résume à la jointure entre les 2 lèvres. Lorsque la bouche est ouverte, le contour interne se décomposera en un contour interne supérieur et un contour interne inférieur. Pour modéliser de manière robuste le contour interne de la bouche il faudra différencier les 2 cas de figure. Dans la suite de ce chapitre, nous développerons nos méthodes de segmentation des contours externes et internes des lèvres.

La section 5.2 sera consacrée à la segmentation du contour externe de la bouche. Nous présenterons nos hypothèses et notre méthodologie. Les étapes de l'algorithme de modélisation du contour seront par la suite développées. La section 5.3 sera consacrée à la segmentation du contour interne des lèvres. Deux méthodes distinctes ont été développées pour les cas des bouches ouvertes et des bouches fermées. Le choix de la méthode repose sur l'état de la bouche que l'on suppose connu (cf. chapitre 3). Enfin, la section 5.4 sera consacrée aux évaluations expérimentales de nos algorithmes de segmentation.

5.2 Segmentation du contour externe des lèvres

Dans le chapitre 1, nous avons présenté un état de l'art des méthodes existantes pour modéliser et segmenter les contours de la bouche. Nous avons pu voir que les 2 grandes

familles de méthodes avaient des limitations différentes. Une approche « région » permettra de localiser la zone de la bouche, mais produira des contours non régularisés et peu précis (voir chapitre 2) que ce soit pour des algorithmes supervisés ou non-supervisés. Les approches « contour » basées sur des snakes ou des modèles paramétriques permettent d'obtenir des modélisations fines des contours mais l'initialisation et le choix des modèles de contour sont très délicats. Pour être performante, une approche utilisant un snake se doit d'initialiser la courbe au plus près du contour à modéliser pour éviter que la courbe converge vers un contour parasite (figure 5.1-a). De plus, la robustesse des snakes aux changements des conditions de l'environnement est très incertaine à la vue du nombre de paramètres à définir (figure 5.1-b). Un modèle paramétrique peut résoudre ce genre de problème mais sa définition est aussi problématique. Trop simple il ne permet pas une modélisation fine de la bouche (figure 5.1-c et 5.1-d) et s'il est trop complexe le temps de convergence peu devenir très important.



Figure 5.1 : Exemples d'échecs des méthodes de modélisation de la bouche par approche contour : a) Exemple d'échec dû à une mauvaise initialisation, à gauche, on donne le contour initial, à droite, on donne le contour final (Delmas, 2000), b) Exemples de convergence d'un snake avec des paramètres différents (Delmas, 2000), c) et d) Exemples de description du contour de la bouche avec un modèle paramétrique trop simple, les croix représentent le contour réel et les lignes blanches représentent le modèle à 2 paraboles après optimisation.

5.2.1 Méthodologie

Dans cette étude, nous faisons l'hypothèse que les lèvres ont été localisées et que le masque des lèvres a été déterminé (voir chapitre 2). Le problème est d'extraire le contour externe de la bouche. Nous avons vu que, pour résoudre ce problème de manière précise, une approche contour est pertinente. Dans notre approche, la recherche du contour externe a été divisée en 4 étapes : modélisation du contour externe supérieur, modélisation du contour externe inférieur, recherche des commissures et modélisation finale du contour (figure 5.2). Pour modéliser les contours externes supérieur et inférieur, nous avons choisi une approche ascendante où les contours seront modélisés par des courbes polynomiales ouvertes.

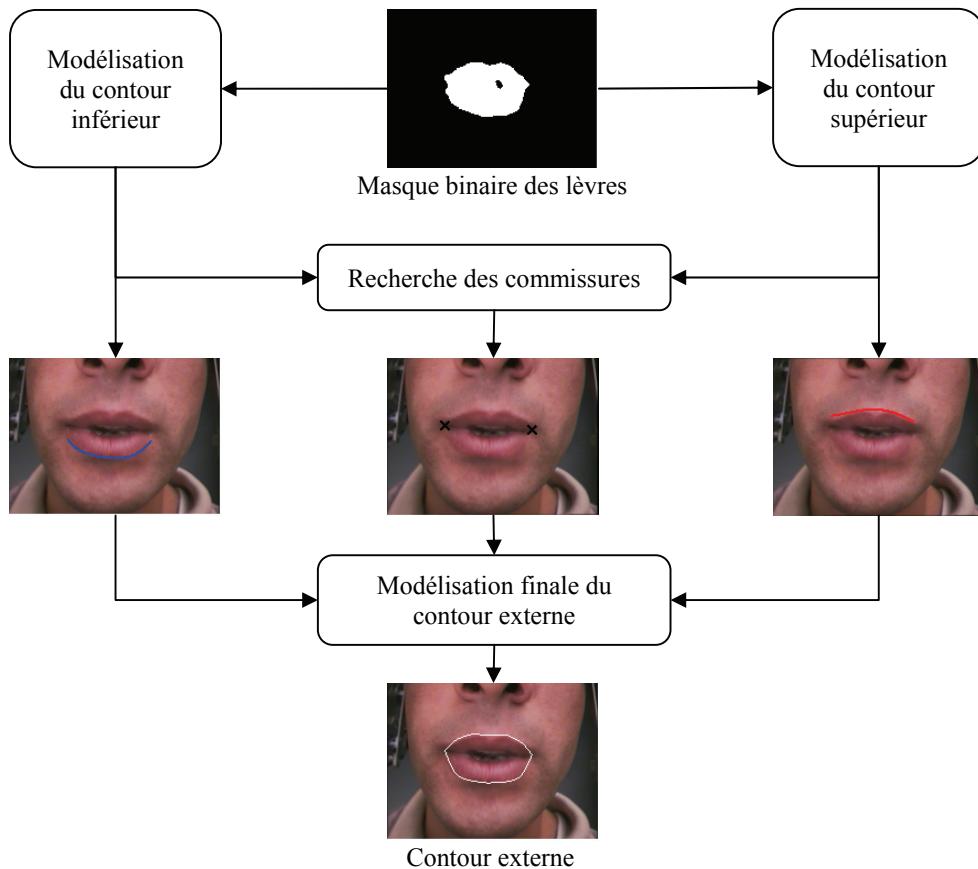


Figure 5.2 : Schéma bloc des étapes de modélisation du contour externe de la bouche à partir du masque binaire des lèvres.

Le masque des lèvres permet de localiser les zones des contours externes supérieur et inférieur. Des contours externes initiaux seront calculés à partir du masque des lèvres. La

force externe $F_{externe}$ est constituée par le flux des gradients γ -normalisés. Pour chaque cas, le degré du polynôme modélisant le contour ainsi que sa position seront optimisés itérativement afin d'ajuster la complexité de la courbe au cas traité. La force interne $F_{interne}$ va s'adapter avec le degré du polynôme. Plus le degré du polynôme est bas, plus les contraintes locales seront importantes et inversement. La recherche des commissures se fera à partir d'un ensemble de positions candidates aux voisinages des extrémités des contours externes supérieur et inférieur. Enfin, à partir des commissures et des contours optimisés, une dernière optimisation permettra d'extraire le contour final.

5.2.2 Optimisation des contours externe supérieur et externe inférieur

Nous supposons, à partir de maintenant, que l'on dispose d'un masque délimitant la zone des lèvres. On restreint alors la zone de recherche à la boîte englobant le masque. Le contour externe du masque des lèvres trouvé précédemment est considéré comme le contour externe initial. En pratique, ce contour se révèle la plupart du temps peu précis. Une étape spécifique de modélisation du contour externe de la bouche est requise. L'intérêt du masque des lèvres est qu'il nous permet d'estimer les proportions de la bouche ainsi que des contours externes supérieur et inférieur initiaux proches des contours des lèvres. Nous avons privilégié une approche « contour » pour la modélisation finale du contour externe de la bouche. Dans la chapitre I, nous avons mis en évidence les problèmes rencontrés avec les méthodes contours classiques comme les snakes ou les méthodes à base de modèles paramétriques. Ces modèles nécessitent la définition de nombreux paramètres qui les rendent sensibles aux changements des conditions de l'environnement. Dans le cas d'un modèle paramétrique il faut définir quel type de courbe sera utilisé pour modéliser les contours suivant les contraintes que l'on fixe sur la forme du modèle. Dans notre cas, le masque des lèvres nous permet de disposer d'un contour externe initial de la bouche. Nous avons choisi de modéliser les contours externes supérieur et inférieur initiaux par 2 courbes polynomiales. La procédure pour optimiser les 2 contours, externe supérieur et externe inférieur, étant identique, nous développerons dans la suite de cette section la procédure pour le cas de l'optimisation du contour supérieur avant de la généraliser pour le contour externe inférieur.

Pour optimiser le modèle de contour externe supérieur sur le contour recherché, sachant que l'on dispose d'un contour initial proche du contour réel, notre idée est d'optimiser une courbe polynomiale en partant d'un modèle simple (une parabole) et d'augmenter itérativement la complexité (le degré du polynôme) tant que la déformation du contour est suffisante et fait croître un critère de performance.

Les étapes de l'optimisation du contour externe supérieur sont les suivantes :

- On extrait la partie P_{sup} supérieure du polygone convexe entourant le masque des lèvres.

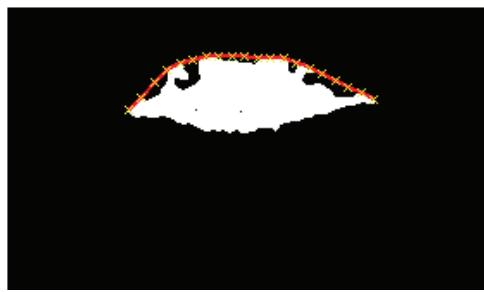


Figure 5.3 : Exemple de contour externe supérieur initial. En rouge, nous avons tracé la partie supérieure du polygone convexe entourant le masque des lèvres que nous considérons comme le contour initial C_{sup} . Les croix jaunes correspondent aux N_{pt} points $KP_{sup} = [K_{sup_1}, \dots, K_{sup_{N_{pt}}}]$ de contrôle.

- On extrait N_{pt} points de contrôle P_{sup} repartis régulièrement sur la largeur de la bouche.
 - On initialise le degré deg_{sup} de la courbe polynomiale modélisant le contour externe supérieur à $deg_{sup} = 2$.
 - La courbe $C_{sup}(deg_{sup})$ est calculée par la méthode des moindres carrés à partir des N_{pt} points de contrôle.
 - La courbe $C_{sup}(deg_{sup})$ est optimisée par déplacements successifs des points de contrôle dans le but de maximiser de la somme des flux des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ à travers la courbe correspondante tant que la déformation de la courbe est suffisante.
 - Lorsque la courbe se stabilise, c'est-à-dire si la norme entre la courbe $C_{sup}(deg_{sup})$ initiale et la courbe optimisée $C_{opt, sup}(deg_{sup})$ est inférieure à un seuil ε_{stop} , on pose $deg_{sup}=deg_{sup}+1$. On répète alors les opérations d'optimisation de C_{sup} avec cette nouvelle valeur deg_{sup} .

- De la même manière, on continue d'incrémenter le degré deg_{sup} de la courbe modélisant le contour externe supérieur tant que la déformation de $Copt_{sup}(deg_{sup})$ par rapport à $Copt_{sup}(deg_{sup}-1)$ est suffisante, c'est-à-dire tant que $|Copt_{sup}(deg_{sup}) - Copt_{sup}(deg_{sup}-1)| > \varepsilon_{stop}$.

Pour réduire le temps de calcul, on échantillonne la partie supérieure du polygone convexe avec N_{pt} points $KP_{sup} = [Ksup_1(x,y), \dots, Ksup_{Npt}(x,y)]$ qui correspondent aux points de contrôle (Figure 5.3) du contour supérieur. Le degré initial deg_{sup} de la courbe polynomiale est fixé à deux. On calcule alors la courbe $C_{sup}(deg_{sup})$ initiale. $C_{sup}(deg_{sup})$ est obtenue par la méthode des moindres carrés à partir des points KP_{sup} . On va ensuite optimiser $C_{sup}(deg_{sup})$ par déformations successives. La déformation de $C_{sup}(deg_{sup})$ est guidée par les déplacements verticaux des points de contrôle $KP_{sup} = [Ksup_1(x,y), \dots, Ksup_{Npt}(x,y)]$. Le but est de trouver la courbe qui maximise la somme des flux des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ pour deg_{sup} .

La déformation de la courbe s'opère comme il suit : on cherche le point de contrôle $K_i(x,y)$ pour lequel la somme des énergies $SE = \sum_{t=1}^{N_{echelle}} |\nabla_{norm}L(P_{min}, t)|^2$ est minimale. On calcule les courbes candidates en faisant varier la position verticale de $K_i(x,y)$ de Δ pixels avec $-LB/4 < \Delta < LB/4$, LB correspond à la hauteur du masque de la bouche. La Figure 5.4 présente l'exemple d'un ensemble de courbes candidates obtenues par déplacement d'un point de contrôle. La nouvelle position $K_{imax}(x,y)$ correspond à celle pour laquelle la somme des flux $FL(C_{sup}(deg_{sup}))$ (49) des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ à travers la courbe C_{sup} est maximale. On appelle cette courbe optimisée $Copt_{sup}(deg_{sup})$.

$$FL(C_{sup}(deg_{sup})) = \sum_{t=1}^{N_{echelle}} \int_{C_{sup}(deg_{sup})} \nabla_{norm}L(x, y, t) dn \quad (49)$$

avec dn le vecteur orthogonal à la courbe, $t=\sigma^2$ l'échelle et (x,y) les coordonnées dans l'image. Nous avons vu que dans \hat{U} , les niveaux des pixels des lèvres étaient inférieurs à ceux des pixels de la peau. L'intégrale sur le contour $C_{sup}(deg_{sup})$ est calculée de manière à ce que la somme soit positive pour la transition peau/lèvres. Dans le cas de la Figure 5.4,

l'intégrale est calculée de la gauche vers la droite de la bouche pour le contour externe supérieur et de la droite vers la gauche pour le contour inférieur.

On observe sur la Figure 5.4 la déformation de la courbe. Dans cet exemple le pas sur Δ a été fixé à 4 pour permettre d'observer la déformation avec un minimum de tracé.

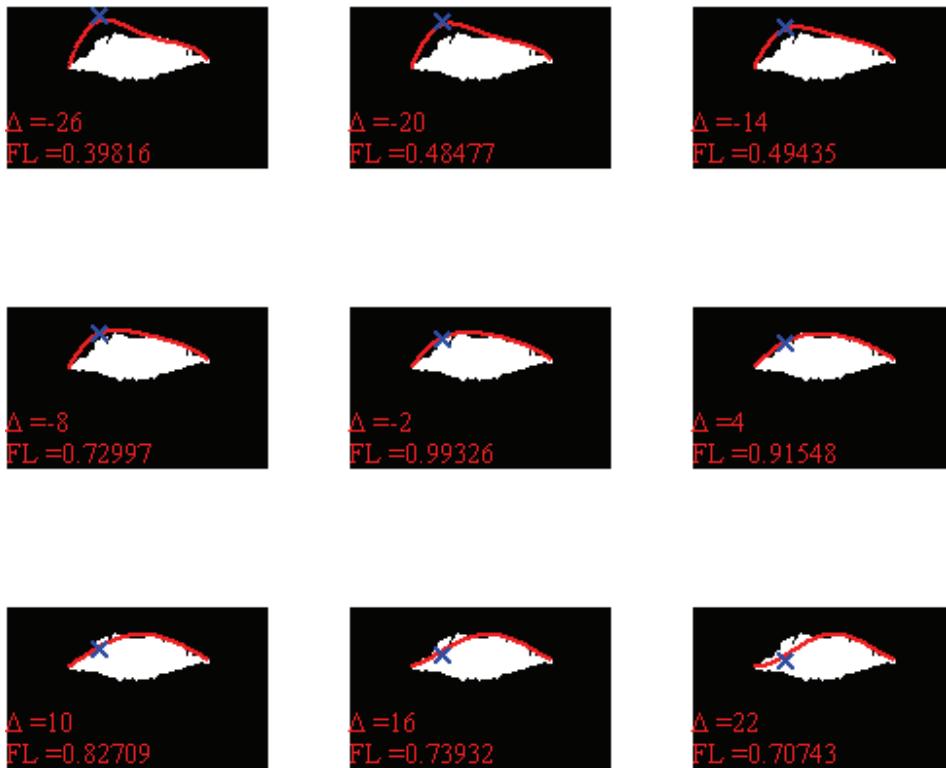


Figure 5.4 : Déformation de la courbe modélisant le contour C_{sup} avec $deg_{sup}=4$ sous l'action du déplacement d'un point de contrôle symbolisé par une croix bleue. On donne la somme FL des flux des gradients γ -normalisés $\nabla_{norm}L(x, y, t)$ à travers la courbe ainsi que le déplacement Δ par rapport à la position verticale d'origine du point de contrôle.

Ensuite, si la déformation de la courbe a été suffisamment importante, précisément si la norme $|C_{sup}(deg_{sup}) - Copt_{sup}(deg_{sup})|$ est supérieure à ε_{stop} , alors on pose $C_{sup}(deg_{sup}) = Copt_{sup}(deg_{sup})$ et on relance l'optimisation par déplacement d'un point de contrôle ; ε_{stop} correspond à 10% de la largeur LB de la bouche. Lorsque $|Copt_{sup}(deg_{sup}) - C_{sup}(deg_{sup})| < \varepsilon_{stop}$, on incrémente $deg_{sup} = deg_{sup} + 1$ et on répète les opérations d'optimisation de $C_{sup}(deg_{sup})$ avec la nouvelle valeur deg_{sup} tant que $|Copt_{sup}(deg_{sup}) - Copt_{sup}(deg_{sup}-1)| > \varepsilon_{stop}$. En pratique l'augmentation du degré du polynôme va autoriser la courbe modélisant le contour externe supérieur à se déformer localement (F_{int} diminue) de manière

plus importante. Lorsque l'augmentation de la complexité du modèle n'entraîne plus de déformation importante de la courbe, on considère que le contour optimal a été trouvé.

On donne à la Figure 5.5 les courbes obtenues avec notre procédure d'optimisation pour des valeurs croissantes de deg_{sup} . L'intérêt de cette procédure d'optimisation est de nous permettre d'adapter la complexité de la courbe modélisant le contour au problème. On observe bien qu'à mesure que le degré de la courbe polynomiale augmente, la modélisation du contour externe supérieur devient de plus en plus précise. Dans ce cas l'optimisation s'est arrêtée à $deg_{sup} = 7$. D'une manière générale, lorsque le degré du polynôme est faible, les contraintes locales sur la courbe sont fortes et limitent la déformation du contour, le modèle de contour est moins sensible au bruit et donnera une modélisation globale du contour. Plus le degré du polynôme sera élevé, plus le contour pourra se déformer localement et plus les détails du contour seront modélisés. Le but de la procédure est d'augmenter le niveau de détail itérativement.

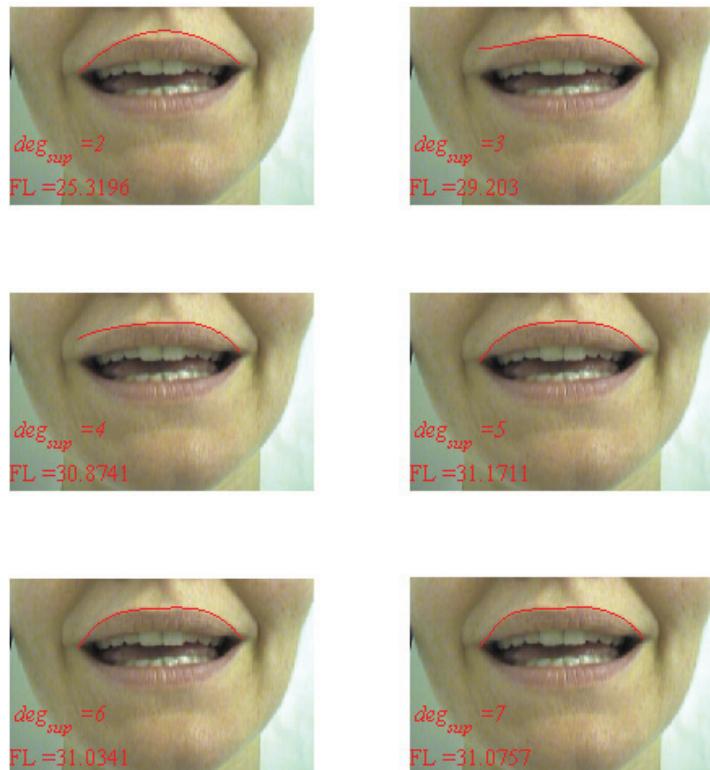


Figure 5.5 : Exemples de courbes C_{sup} obtenues avec notre procédure d'optimisation pour des valeurs croissantes de deg_{sup} pour une image de bouche ouverte.

La procédure d'optimisation est identique pour la modélisation du contour externe inférieur C_{inf} . Le modèle de contour C_{inf} est initialisé sur la partie inférieure du polygone convexe entourant le masque de la zone des lèvres (Figure 5.6). Après déformations successives du contour initial C_{inf} , nous obtenons une courbe modélisant le contour externe inférieur de la bouche, comme l'exemple de la Figure 5.7. Nous avons tracé la courbe C_{inf} obtenue en rouge ainsi que les positions des points de contrôle obtenues. On donne également le degré deg_{inf} de la courbe et la valeur du flux FL pour cette courbe.

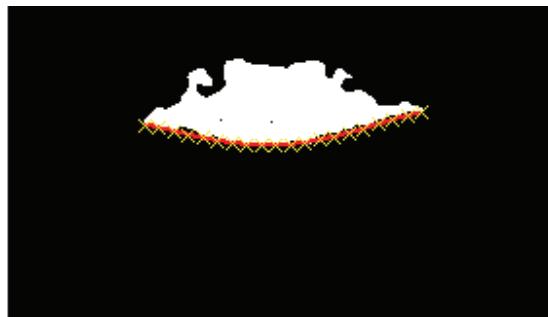


Figure 5.6 : Exemple de contour externe inférieur initial. En rouge, nous avons tracé la partie inférieure du polygone convexe que nous considérons comme le contour initial. Les croix jaunes correspondent aux N_{pt} points $KP_{inf} = [K_{inf_1}, \dots, K_{inf_{Npt}}]$ de contrôle.



Figure 5.7 : Exemple de contour externe inférieur C_{inf} obtenu après optimisation par notre méthode. En rouge nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.

Le choix du nombre de points de contrôle N_{pt} va influencer la vitesse et la précision de la modélisation du contour. Tout d'abord, le choix de N_{pt} va limiter la complexité de la courbe. Pour pouvoir calculer la courbe polynomiale, il faut que $deg_{sup} \leq N_{pt}$. De plus, comme les courbes C_{sup} et C_{inf} sont estimées par la méthode des moindres carrés, il est nécessaire que le nombre de points de contrôle soit supérieur au degré de la courbe pour éviter le problème de mauvais conditionnement. Pour avoir une description suffisamment

précise du contour externe de la bouche, il faudra donc utiliser un nombre minimum de points de contrôle. Expérimentalement nous avons utilisé $N_{pt} = 20$ points répartis régulièrement pour modéliser avec une précision suffisante les contours externe supérieur et externe inférieur.

On donne sur la Figure 5.8 un exemple d'optimisation des contours externe supérieur et externe inférieur pour des images de bouche. Le contour externe supérieur C_{sup} est tracé en rouge et le contour externe inférieur C_{inf} en bleu.



Figure 5.8 : Exemples d'optimisation des contours externe supérieur et externe inférieur pour des images de bouche.

5.2.3 Recherche des commissures

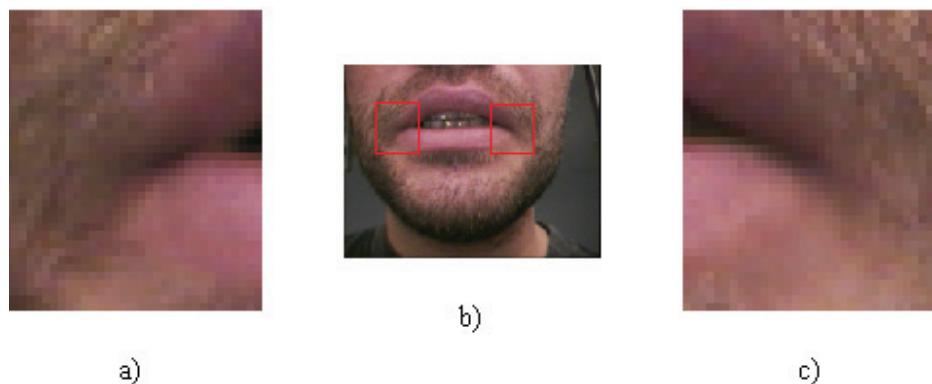


Figure 5.9 : Zoom sur les zones des commissures de la bouche, a) zone de la commissure gauche, b) image de bouche, c) zone de la commissure droite.

A partir de maintenant, on suppose que l'on dispose des contours C_{sup} et C_{inf} optimisés par notre méthode. Précisément on dispose des ensembles de points de contrôle $KP_{sup} = [Ksup_1(x,y), \dots, Ksup_{Npt}(x,y)]$ et $KP_{inf} = [Kin_1(x,y), \dots, Kin_{Npt}(x,y)]$ ainsi que les degrés deg_{sup} et deg_{inf} optimaux permettant le calcul de C_{sup} et C_{inf} . L'étape suivante est la recherche des commissures pour fermer le contour externe de la bouche. Les commissures

de la bouche représentent les points de jonction des contours C_{sup} et C_{inf} . La localisation des commissures de la bouche est un problème difficile. La plupart du temps, il est impossible de définir ces points par des caractéristiques locales comme la teinte ou le gradient. La Figure 5.9 illustre la difficulté de localiser les commissures. En effectuant un zoom sur la zone où se trouve subjectivement la commissure droite, on remarque à quel point il est difficile de la placer visuellement. Les contours de la bouche sont mal définis et la zone où se trouve a priori la commissure est sombre. Un expert humain localisera implicitement les commissures par rapport à la forme globale de la bouche en suivant grossièrement les contours C_{sup} et C_{inf} jusqu'aux points où ils se rejoignent pour trouver les commissures. Notre approche pour localiser les commissures s'inspire de ce principe. Les contours C_{sup} et C_{inf} nous donnent une modélisation du contour externe de la bouche et le masque des lèvres permet d'identifier les zones dans lesquelles se trouvent a priori les commissures. Eveno propose d'utiliser une méthode de chaînage pour localiser les commissures (Eveno, 2003). Un germe est initialisé sur la colonne centrale de la bouche. A partir du germe, la ligne L_{min} est étendue à droite au point le plus sombre parmi les 3 voisins directs du germe, et ainsi de suite jusqu'à la limite de l'image. La ligne est étendue de la même manière à gauche. Eveno émet alors l'hypothèse que les commissures appartiennent à la ligne L_{min} .

Dans notre cas, les contours optimisés C_{sup} et C_{inf} nous permettent de réduire la zone de recherche des commissures par rapport à Eveno. A priori les extrémités gauche et droite de C_{sup} et C_{inf} se trouvent au voisinage des commissures. Nous émettons une hypothèse analogue à celle de (Delmas, 2000) et (Eveno, 2003) qui est que les commissures correspondent à des zones de faible luminance et proches des lèvres. Les contours vont nous permettre d'initialiser 2 germes, un germe pour la commissure gauche et un germe pour la commissure droite. Ces 2 germes serviront alors à construire 2 lignes Lg_{min} et Ld_{min} chaînant les pixels sombres, respectivement, aux voisinages des commissures gauche et droite. Le germe $Gg(xg,yg)$ de Lg_{min} est initialisé en cherchant le pixel le plus sombre aux extrémités gauches de C_{sup} et C_{inf} avec la contrainte que celui-ci soit à l'intérieur de la zone de la bouche. En pratique nous avons utilisé un décalage de 1 % de la largeur de la zone de la bouche vers l'intérieur. À partir de $Gg(xg,yg)$, la ligne Lg_{min} est étendue vers la gauche au voisin de plus faible luminance, parmi les 3 voisins directs du germe. L'opération de chaînage est répétée jusqu'à la limite gauche de l'image de bouche. La Figure 5.10 présente

le tracé de Lg_{min} à parti du germe Gg . Pour la commissure droite, la méthode est identique par symétrie. $Gd(xd,yd)$ est initialisé au pixel le plus sombre entre les contours C_{sup} et C_{inf} au voisinage de l'extrémité droite de la bouche et Ld_{min} est étendue vers la droite (Figure 5.10).

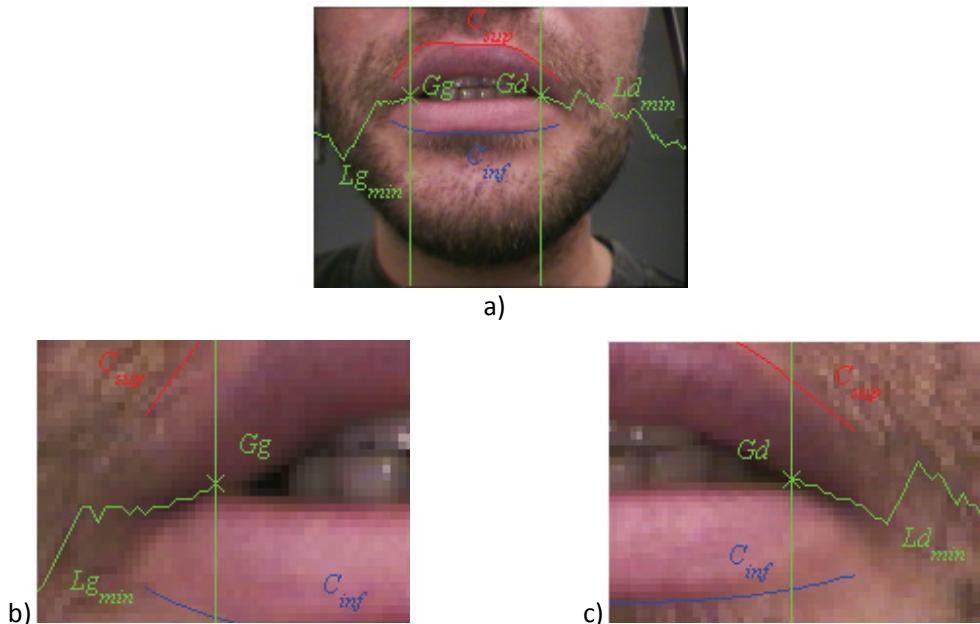


Figure 5.10 : Tracés des lignes Lg_{min} et Ld_{min} pour une image de bouche, a) Image d'entrée, b) Zoom sur la zone de la commissure gauche, c) zoom sur la zone de la commissure droite.

Si nous faisons l'hypothèse que les commissures gauche et droite appartiennent respectivement à Lg_{min} et Ld_{min} , il reste alors à tester les points de Lg_{min} et Ld_{min} comme points de jonction de C_{sup} et C_{inf} , et à conserver le couple de points pour lequel notre fonction de coût est maximisée. Dans notre cas, la fonction de coût correspond à la somme FL des flux à travers le contour pour l'ensemble des échelles considérées. Dans la pratique, la recherche des commissures peut être réalisée séparément. Nous développerons dans la suite le cas de la commissure gauche.

Pour chaque point candidat PLg_{min} on calcule le meilleur couple de courbes $\{C'_{sup}, C'_{inf}\}$ au sens de notre fonction de coût de la manière suivante :

- On pose $m = 1$.
- On construit les ensembles de points $KP'_{sup} = [PLg_{min}(x,y), Ksup_m(x,y), \dots, Ksup_{Npt}(x,y)]$ et $KP'_{inf} = [PLg_{min}(x,y), Kinf_m(x,y), \dots, Kinf_{Npt}(x,y)]$.

- On calcule les courbes $\{C'_{sup}, C'_{inf}\}$ par la méthode des moindres carrés à partir de KP'_{sup} et KP'_{inf} en imposant la contrainte que les courbes passent par PLg_{min} .
- On calcule la somme des flux FL à travers les courbes $\{C'_{sup}, C'_{inf}\}$ pour les échelles considérées :

$$FL = \sum_{t=1}^{N_{echelle}} \int_{\{C'_{sup}, C'_{inf}\}} \nabla_{norm} L(x, y, t) dn \quad (50)$$

- On répète les étapes précédentes tant que $m < N_{pt}$.
- Le couple optimal pour PLg_{min} est celui pour lequel FL est maximale et l'indice m correspondant est appelé mg_{opt} .

La Figure 5.11 illustre les étapes exposées précédemment. Nous avons souligné que les zones aux voisinages des commissures étaient souvent sombres et avec des contours externes peu marqués. Les extrémités des contours C_{sup} et C_{inf} peuvent être bruitées. Un couple de courbes $\{C'_{sup}, C'_{inf}\}$ est calculé par moindres carrés à partir des ensembles de points $KP'_{sup} = [PLg_{min}(x,y), Ksup_m(x,y), \dots, Ksup_{Npt}(x,y)]$ et $KP'_{inf} = [PLg_{min}(x,y), Kinf_m(x,y), \dots, Kinf_{Npt}(x,y)]$ et des degrés deg_{sup} et deg_{inf} obtenus à l'étape précédente.

En augmentant l'indice m , nous allons progressivement éliminer de l'estimation des courbes $\{C'_{sup}, C'_{inf}\}$ les points de contrôle issue de KP_{sup} et KP_{inf} et se trouvant aux extrémités de la bouche. Sur la Figure 5.11, on observe que la position de la commissure candidate est fixe et qu'itérativement on retire les points de contrôle issue des ensemble KP_{sup} et KP_{inf} pour calculer $\{C'_{sup}, C'_{inf}\}$. De cette manière, la pente des courbes $\{C'_{sup}, C'_{inf}\}$ au point de jonction PLg_{min} va varier jusqu'à obtenir le couple $\{C'_{sup}, C'_{inf}\}$ optimal pour cette commissure candidate. Pour chaque point de Lg_{min} , nous disposerons d'un couple $\{C'_{sup}, C'_{inf}\}$ calculé à partir des ensembles $KG'_{sup} = [PLg_{min}, Ksup_{mgopt}, \dots, Ksup_{Npt}]$ et $KG'_{inf} = [PLg_{min}, Kinf_{mgopt}, \dots, Kinf_{Npt}]$ qui maximise FL (Figure 5.10-a). Le point PLg_{opt} de Lg_{min} , pour lequel FL est maximum, est considéré comme la commissure gauche de la bouche (Figure 5.12-a). On nomme les ensembles de points permettant le

calcul du couple de courbes associé à ce point de la manière suivante : $KGopt_{sup} = [PLgopt_{min}, Ksup_{mg}, \dots, Ksup_{Npt}]$ et $KGopt_{inf} = [PLgopt_{min}, Kinf_{mg}, \dots, Kinf_{Npt}]$ avec $mg = mg_{opt}$. La même méthode est employée de manière symétrique pour trouver la commissure droite de la bouche. Nous obtenons les ensembles de points de contrôle suivant : $KDopt_{sup} = [Ksup_1, \dots, Ksup_{md}, PLdopt_{min}]$ et $KDopt_{inf} = [Kin_1, \dots, Kin_{md}, PLdopt_{min}]$ avec $md = md_{opt}$.

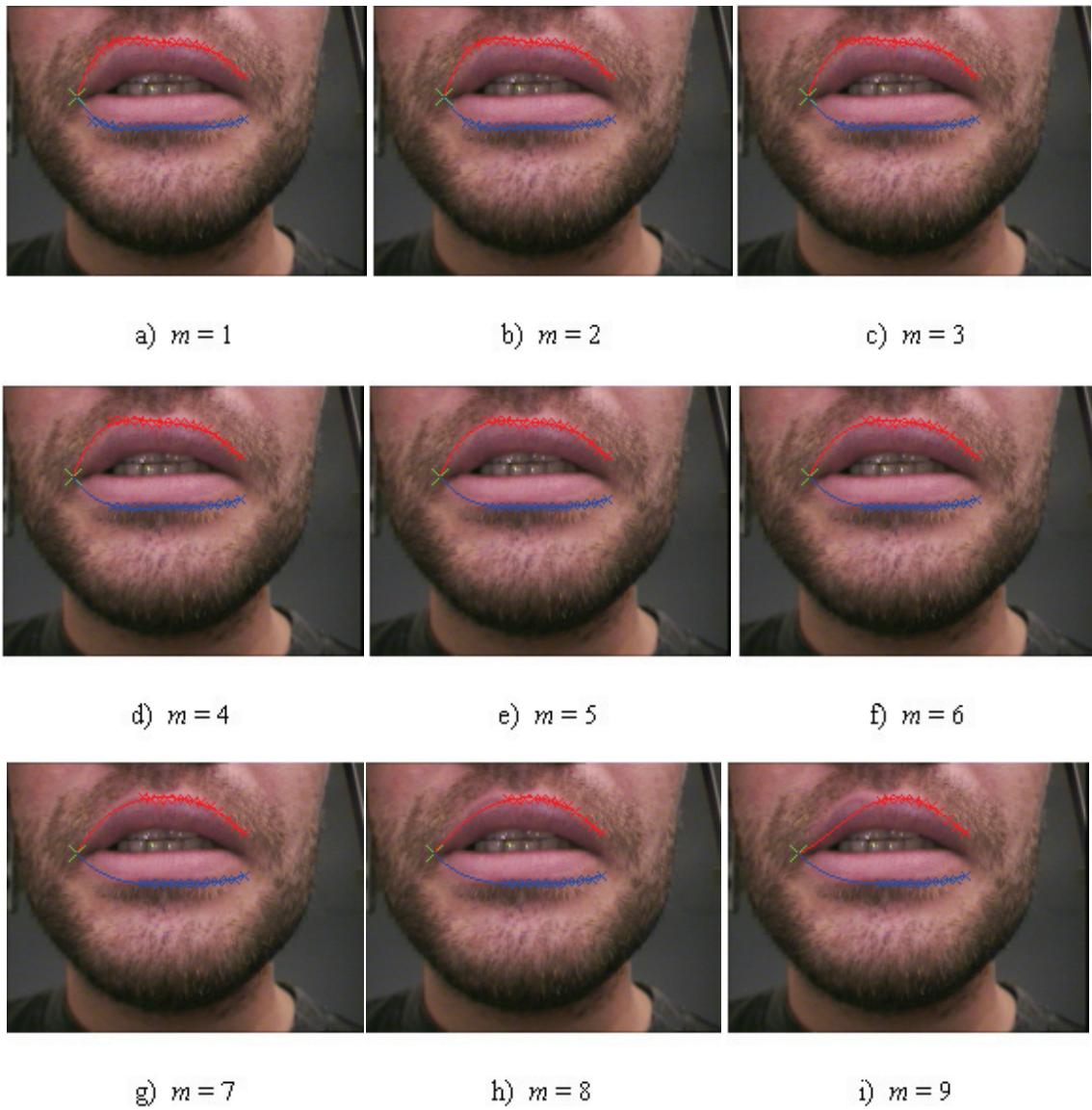


Figure 5.11 : Tracés des déformations des courbes $\{C'_{sup}, C'_{inf}\}$ lorsque l'indice m augmente pour un point PLg_{min} de Lg_{min} particulier. Les tracés en trait plein correspondent aux courbes $\{C'_{sup}, C'_{inf}\}$ calculées par la méthode des moindres carrés à partir des points de contrôle marqués par des croix.

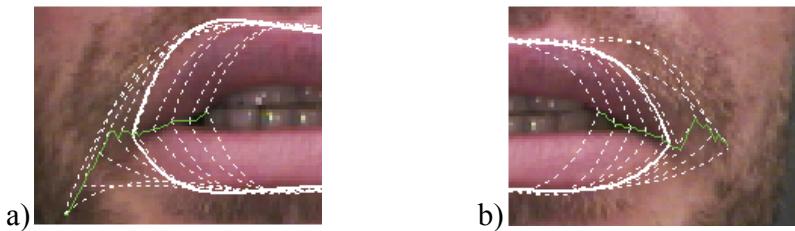


Figure 5.12 : Tracés des couples $\{C'_{sup}, C'_{inf}\}$ pour différents points de Lg_{min} et Ld_{min} . En trait plein blanc, on donne les couples maximisant FL . En pointillés, on donne les tracés des couples candidats, et en vert les tracés de Lg_{min} et Ld_{min} , a) cas de la commissure gauche, b) cas de la commissure droite.

5.2.4 Modélisation finale du contour externe de la bouche

Nous disposons maintenant de suffisamment d'information pour modéliser complètement le contour externe de la bouche. Nous disposons des ensembles de points de contrôle $KGopt_{sup}$, $KGopt_{inf}$, $KDopt_{sup}$ et $KDopt_{inf}$. À partir des points de ces ensembles, nous construisons alors les ensembles de points de contrôle suivants : $K_{sup} = [PLgopt_{min}, Ksup_{mg}, \dots, Ksup_{md}, PLdopt_{min}]$ et $K_{inf} = [PLgopt_{min}, Kinf_{mg}, \dots, Kinf_{md}, PLdopt_{min}]$. À partir de ces ensembles, on répète alors les étapes d'optimisation décrites à la section 5.2.2 pour ajuster la complexité des modèles de courbes aux nouveaux ensembles de points de contrôle. Ces ensembles nous permettent de calculer les contours, externe supérieur et externe inférieur, finaux par la méthode des moindres carrés avec les contraintes que les courbes se rejoignent en $PLgopt_{min}$ et $PLdopt_{min}$. La Figure 5.13 présente le contour externe final extrait à partir de l'image de bouche de la Figure 5.9.

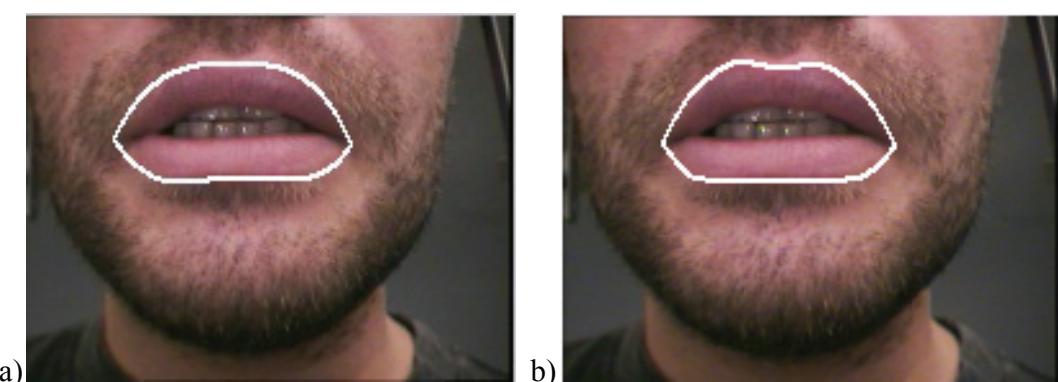


Figure 5.13 : Exemple de modélisation du contour externe de la bouche, a) avant l'étape d'optimisation finale, b) après optimisation finale.

5.3 Segmentation du contour intérieur de la bouche

La configuration de la région interne de la bouche peut être très variable et rendre difficile la modélisation des contours internes. Lorsque celle-ci est fermée, le contour interne se résume à la jointure entre les 2 lèvres. Lorsque la bouche est ouverte, il y aura, comme pour le contour externe, un contour interne supérieur et un contour interne inférieur. Nous avons donc choisi de distinguer les cas des bouches ouvertes et fermées. Au chapitre 3, nous avons proposé une méthode fréquentielle de détection de l'état ouvert/fermé de la bouche pour identifier les 2 cas. Nous détaillerons, à la section 5.3.1, la modélisation du contour interne pour le cas des bouches ouvertes avant de développer la méthode de modélisation du contour interne pour les bouches fermées.

5.3.1 Modélisation du contour interne pour les bouches ouvertes

5.3.1.1 Méthodologie

Dans cette étude, nous faisons l'hypothèse que le contour externe de la bouche a été segmenté dans son ensemble et que la bouche est identifiée comme ouverte (voir le chapitre 3 et section 5.2). Le problème est, maintenant, d'extraire le contour interne. Dans un premier temps, nous nous sommes intéressés aux grandeurs colorimétriques qui offrent la meilleure séparation entre les lèvres et la bouche. Aux sections 1.2 et 2.3, nous avons étudié le pouvoir de séparation entre la peau et les lèvres de différentes grandeurs colorimétriques pour déterminer quelle grandeur est la mieux adaptée à la modélisation des lèvres. Nous avons montré que l'algorithme allongement-décorrélation permettait d'augmenter de manière importante le contraste peau/lèvres et en particulier pour la grandeur \hat{U} que nous avons choisie pour localiser la bouche et pour modéliser le contour externe de la bouche. Nous avons répété l'étude menée à la section 2.3, mais en nous intéressant cette fois à la séparation entre les lèvres et la zone interne de la bouche. Le but est de chercher la ou les grandeurs offrant le meilleur contraste entre les lèvres et la zone interne de la bouche. À partir de notre base de test de 150 images de 20 sujets différents, nous avons extrait les pixels des lèvres et de la zone interne de la bouche manuellement sur les bouches ouvertes. Nous avons ensuite recalculé les variances intraclasses et interclasses ainsi que les rapports V_{intra}/V_{inter} pour les grandeurs colorimétriques, étudiées au chapitre 1, calculées à partir de

R_{decorr} , G_{decorr} , B_{decorr} normalisées au préalable entre 0 et 1. Les résultats sont donnés à la table 5.1. À la figure 5.14, nous avons tracé les histogrammes des distributions des ensembles de pixels des lèvres et des pixels appartenant à la zone interne de la bouche.

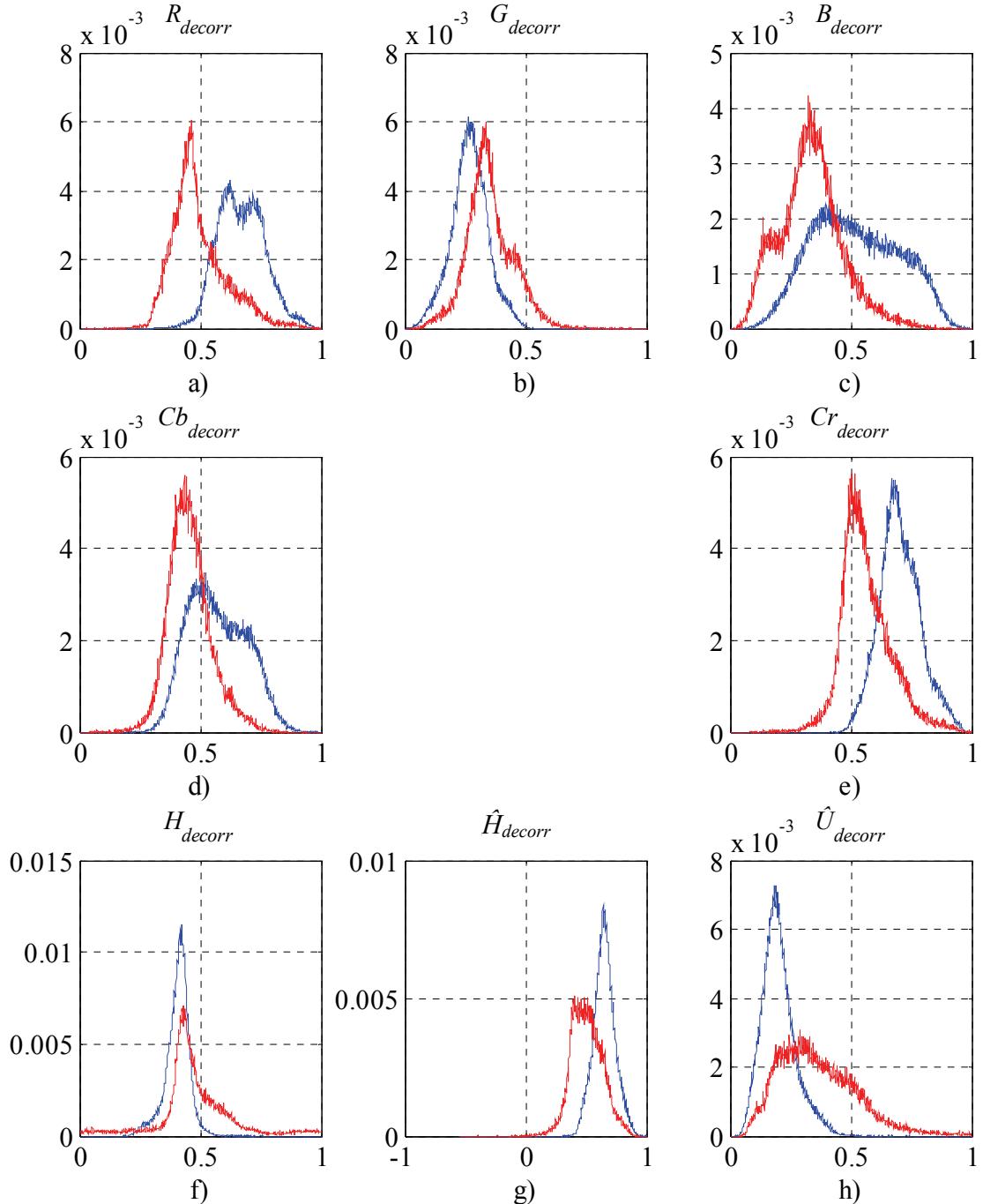


Figure 5.14 : Tracés des histogrammes des distributions des ensembles de pixels de la zone interne de la bouche (en rouge) et des lèvres (en bleu), a) R_{decorr} , b) G_{decorr} , c) B_{decorr} , d) Cb_{decorr} , e) Cr_{decorr} , f) H_{decorr} , g) \hat{H}_{decorr} et h) \hat{U}_{decorr} .

	Variance intraclasses	Variance interclasses	V_{intra}/V_{inter}
R_{decorr}	$2.23 \cdot 10^{-2}$	$1.41 \cdot 10^{-2}$	1.58
G_{decorr}	$1.4 \cdot 10^{-2}$	$2.22 \cdot 10^{-3}$	4.7
B_{decorr}	$2 \cdot 10^{-2}$	$5.3 \cdot 10^{-3}$	3.83
Cb_{decorr}	$4.5 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	4.4
Cr_{decorr}	$1.03 \cdot 10^{-3}$	$5.3 \cdot 10^{-3}$	2.01
H_{decorr}	$9 \cdot 10^{-2}$	$9.95 \cdot 10^{-4}$	9
\hat{H}_{decorr}	$1.92 \cdot 10^{-2}$	$8.3 \cdot 10^{-3}$	2.32
\hat{U}_{decorr}	$4.9 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	1.88

Table 5.1 : Variance intraclasses, variance interclasses et V_{intra}/V_{inter} pour les composantes R_{decorr} , G_{decorr} , B_{decorr} , Cb_{decorr} , Cr_{decorr} , H_{decorr} , \hat{H}_{decorr} et \hat{U}_{decorr} pour le cas de la séparation lèvres/intérieur de la bouche.

Les résultats de la table 5.1 montrent que le recouvrement est important entre les distributions des ensembles des pixels des lèvres et des pixels de la zone interne de la bouche pour toutes les grandeurs colorimétriques que nous avons étudiées. Le rapport V_{intra}/V_{inter} est minimal dans notre étude pour la composante R_{decorr} . La teinte \hat{U}_{decorr} et la composante Cr_{decorr} viennent ensuite. À la figure 5.15, nous donnons les composantes R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} pour une image de bouche ouverte avec la présence des dents et de la langue.

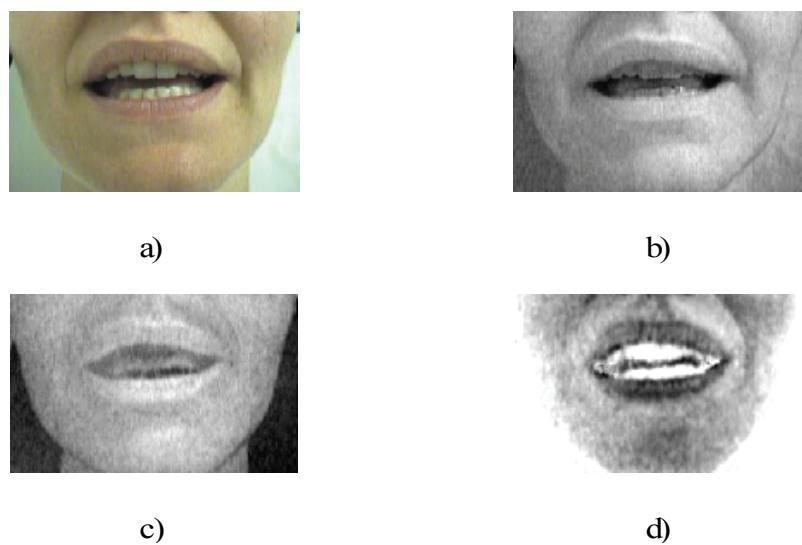


Figure 5.15 : Image de bouche ouverte pour différentes grandeurs chromatiques, a) image RGB , b) R_{decorr} , c) Cr_{decorr} , d) \hat{U}_{decorr} .

On constate que pour les composantes R_{decorr} , Cr_{decorr} , la zone interne de la bouche est plus sombre que les lèvres, que ce soit les dents ou l'intérieur de la bouche. Sur la teinte \hat{U}_{decorr} , les dents présentent un très fort contraste avec le reste de la bouche mais l'intérieur de la bouche, entre les dents, est dans les mêmes gammes de teintes que les lèvres. Sur la figure 5.16, nous avons également donné les images des intensités des gradients pour les 3 composantes R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} . Pour améliorer la visualisation pour chaque composante, l'image d'intensité correspond à la somme des intensités des gradients γ -normalisés des échelles $t=\sigma^2$ avec $\sigma=\{1,2,3\}$.

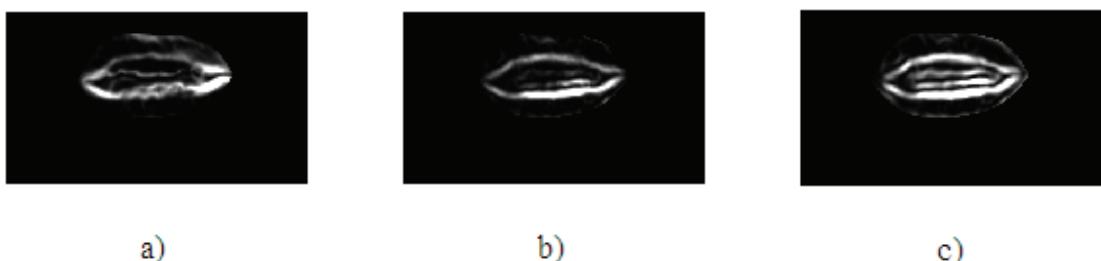


Figure 5.16 : Images des sommes des intensités des gradients γ -normalisés pour les composantes R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} pour les échelles $t=\sigma^2$ avec $\sigma=\{1,2,3\}$, a) sommes des intensités pour R_{decorr} , c) somme des intensités pour Cr_{decorr} , d) somme des intensités pour \hat{U}_{decorr} .

On observe sur la figure 5.16 que les contours internes des lèvres sont bien définis dans les 3 composantes. Néanmoins, il subsiste des contours parasites, notamment entre les dents et l'intérieur de la bouche. Pour sélectionner les contours d'intérêt (les contours internes des lèvres), nous avons choisi d'utiliser les 3 composantes chromatiques R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} . Notre but est de segmenter l'ensemble des régions internes de la bouche sous forme d'un masque binaire. Ce masque binaire permettra d'initialiser des contours internes supérieur et inférieur et de procéder à une modélisation du contour analogue à celle qui a été appliquée pour le contour externe de la bouche (optimisation des contours internes supérieur et inférieur des lèvres, extraction des commissures internes et modélisation finale du contour). La différence par rapport au cas du contour externe sera l'utilisation des gradients γ -normalisés des grandeurs R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} qui se sont révélées pertinentes pour caractériser les contours internes des lèvres.

5.3.1.2 Segmentation des régions internes de la bouche

Nous avons vu à la section précédente que, les grandeurs R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} permettent de séparer les lèvres et les régions internes de la bouche. Nous avons également vu que le recouvrement entre les distributions des ensembles de pixels constitués sur notre base est important. En effet, dans certains cas, la langue ou les gencives, dont les propriétés chromatiques sont proches de celles des lèvres, peuvent être présentes dans la zone interne de la bouche et rendre la segmentation de l'ensemble de la zone interne de la bouche difficile. Pour segmenter les zones internes de la bouche, nous proposons d'appliquer une méthode de segmentation « région-contour » itérative basée sur l'algorithme des K-moyennes (voir la section 2.5.2.2 pour la présentation de l'algorithme des K-moyennes). Soit l'ensemble de pixels $P_{bouche} = [\hat{ub}_1 \dots \hat{ub}_{Nbouche}]$ avec N_{bouche} le nombre de pixels de la zone de la bouche, les \hat{ub}_i sont des vecteurs constitués des niveaux des pixels de la zone de la bouche dans R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} . Soit les représentations multi-échelle $L_{Rdecorr}(x,y,t)$, $L_{Crdecorr}(x,y,t)$ et $L_{\hat{U}decorr}(x,y,t)$, respectivement, des composantes R_{decorr} , Cr_{decorr} et \hat{U}_{decorr} de l'image de bouche. Nous souhaitons partitionner l'ensemble des pixels entourés par le contour externe de la bouche en M_{bouche} classes $\{Y_1 \dots Y_{M_{bouche}}\}$ par l'algorithme des K-moyennes. On se propose, ensuite, d'éliminer successivement les ensembles de pixels se situant sur la périphérie de la zone de la bouche. Ces ensembles correspondent a priori aux lèvres. On cherchera l'ensemble de blobs dont le contour maximisera les flux des gradients γ -normalisés de $L_{Rdecorr}(x,y,t)$, $L_{Crdecorr}(x,y,t)$ et $L_{\hat{U}decorr}(x,y,t)$. Le nombre M_{bouche} pouvant varier suivant la configuration de la bouche, il est initialisé à deux. L'algorithme de segmentation est répété pour des valeurs croissantes de M_{bouche} afin de déterminer le masque optimal. Les étapes de l'algorithme de segmentation des régions internes de la bouche sont les suivantes :

- On détermine la partition rigide Q_{bouche} en M_{bouche} classes $\{Y_1 \dots Y_{M_{bouche}}\}$ de l'ensemble P_{bouche} des pixels de la bouche, où $q_{i,j} = 1$ quand le stimulus \hat{ub}_j appartient à la classe Y_i (figure 5.17-a).

$$Q_{bouche} = \begin{bmatrix} q_{1,1} & \cdots & q_{1,N_{bouche}} \\ \vdots & \ddots & \vdots \\ q_{M_{bouche},1} & \cdots & q_{M_{bouche},N_{bouche}} \end{bmatrix} \quad (51)$$

- Ensuite pour chaque classe Y_i , on effectue un étiquetage de l'ensemble des blobs non-connexes formés par les pixels de cette classe. De cette manière, les blobs non-connexes d'une même classe auront des étiquettes différentes. On obtient l'image $E(M_{bouche})$ de la figure 5.17-b à partir de l'étiquetage de la figure 5.17-a.

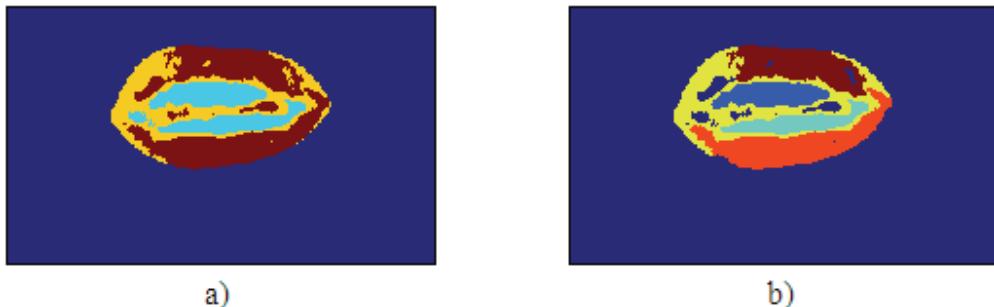


Figure 5.17 : Représentation de la partition Q_{bouche} pour $M_{bouche} = 3$ et de l'image des blobs étiquetés $E(M_{bouche})$, a) Q_{bouche} , b) $E(M_{bouche})$.

On s'intéresse d'abord au contour interne supérieur :

- On élimine ensuite successivement les blobs dont la limite supérieure est la plus haute. On calcule le flux $FLhint$ (52) à travers la partie supérieure du polygone convexe entourant les blobs restants (figure 5.18).



Figure 5.18 : Elimination successive des blobs situés dans la partie supérieure de la bouche et tracés des parties supérieures des polygones convexes entourant les blobs restants.

- On continue tant que $FLhint$ est supérieur ou égal à la valeur précédente (figure 5.18).
- Lorsque le flux $FLhint$ calculé est inférieur à la valeur précédente, on arrête la recherche du contour interne supérieur et on conserve le dernier ensemble de

blobs. Enfin, on extrait le masque binaire correspondant qu'on appelle E_{haut} (figure 5.19).



Figure 5.19 : Représentation de E_{haut} .

- On s'intéresse ensuite au contour interne inférieur.
- A partir de l'image initiale $E(M_{bouche})$, on élimine successivement les blobs dont la limite inférieure est la plus basse. On calcule également $FLbint$ (52) à travers la partie inférieure du polygone convexe entourant les blobs restants (figure 5.20).



Figure 5.20 : Elimination successive des blobs situés dans la partie inférieure de la bouche et tracés des parties inférieures des polygones convexes entourant les blobs restants.

- On continue tant que $FLbint$ est supérieur ou égal à la valeur précédente (figure 5.20).
- Lorsque le flux $FLbint$ calculé est inférieur à la valeur précédente, on arrête la recherche du contour supérieur et on conserve le dernier ensemble de blobs. On extrait le masque binaire correspondant qu'on appelle E_{bas} (figure 5.21).



Figure 5.21 : Représentation de E_{bas} .

- On applique un « ET » logique entre les masques E_{haut} et E_{bas} pour obtenir le masque binaire de la zone interne la bouche (figure 5.22).



Figure 5.22 : Masque binaire de la zone interne la bouche.

- On incrémente M_{bouche} et on répète l'algorithme de segmentation tant que la somme $FLhint + FLbint$ augmente.

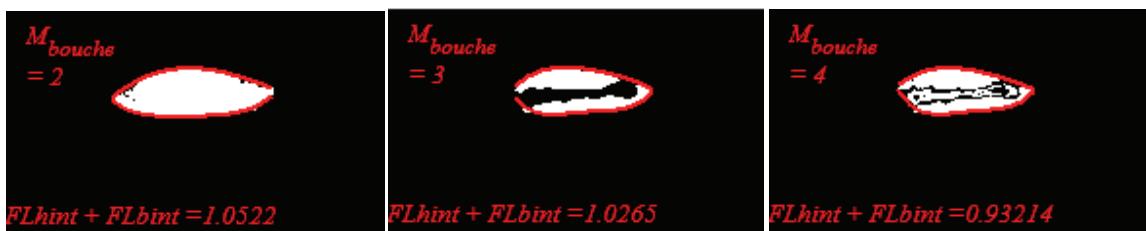


Figure 5.23 : Segmentation de la région interne d'une bouche ouverte pour M_{bouche} croissant.

- Lorsque la somme $FLhint + FLbint$ calculée est inférieure à la valeur précédente, on arrête la segmentation et on conserve le masque binaire précédent. Pour le cas de la figure 5.23, on conserve alors le masque pour $M_{bouche}=2$.

Les expressions des flux $FLhint$ et $FLbint$ à travers les contours internes supérieur C_{hint} et interne inférieur C_{bint} sont données à l'équation (52). L'étude des grandeurs colorimétriques pour la segmentation de l'intérieur de la bouche (cf. figure 5.14) a montré que, dans les grandeurs R_{decorr} et Cr_{decorr} , la zone interne de la bouche, ainsi que les dents, ont des niveaux globalement inférieurs à ceux des lèvres tandis que pour \hat{U}_{decorr} c'est l'inverse. Nous avons pondéré les gradients de $L_{Rdecorr}$ et $L_{Crdecorr}$ par un coefficient -1 dans la somme des gradients pour que les transitions peau/lèvres dans $L_{Rdecorr}$ et $L_{Crdecorr}$ soient du même

signe que dans \hat{U}_{decorr} (52). Le nombre d'échelle $N_{echelle}$ est celui déterminé lors de la segmentation de la bouche sur le visage (cf. section 2.5.3.2).

$$\begin{aligned}
 FLhint &= \sum_{t=1}^{N_{echelle}} \int_{C_{hint}} [\nabla_{norm} L_{\hat{U}_{decorr}}(x, y, t) - \nabla_{norm} L_{R_{decorr}}(x, y, t) \\
 &\quad - \nabla_{norm} L_{Cr_{decorr}}(x, y, t)] dn \\
 FLbint &= \sum_{t=1}^{N_{echelle}} \int_{C_{bint}} [\nabla_{norm} L_{\hat{U}_{decorr}}(x, y, t) - \nabla_{norm} L_{R_{decorr}}(x, y, t) \\
 &\quad - \nabla_{norm} L_{Cr_{decorr}}(x, y, t)] dn
 \end{aligned} \tag{52}$$

5.3.1.3 Modélisation du contour interne des lèvres

Dans la section précédente, nous avons décrit notre méthode pour segmenter la région interne d'une image de bouche ouverte (la zone entre les lèvres). On suppose à partir de maintenant qu'on dispose d'un masque binaire de cette zone. Le but recherché est de segmenter le contour interne des lèvres. Comme pour le contour externe, le masque obtenu a des contours qui la plupart du temps sont bruités, peu précis et il ne permet pas une localisation précise des commissures. Une étape d'optimisation est nécessaire pour modéliser précisément les contours internes, supérieur et inférieur, et localiser les commissures internes. Etant donné la similitude entre la modélisation du contour interne et du contour externe, nous avons appliqué une procédure identique à celle mise en œuvre pour le contour externe, à la différence que nous avons utilisé les sommes $FLhint$ et $FLbint$ (52) comme critères de performance.



Figure 5.24 : Exemple de contour interne supérieur obtenu après optimisation. En rouge, nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.

Une optimisation du contour interne supérieur est d'abord effectuée. La méthode est identique à celle présentée à la section 5.2.2 à la différence que le flux est calculé d'après l'expression (52). On aboutit à un contour modélisé par une courbe polynomiale (figure 5.24). De la même manière que dans le cas du contour externe inférieur, on effectue l'optimisation du contour interne inférieur avec la nouvelle expression du flux comme critère de performance (figure 5.25).



Figure 5.25 : Exemple de contour interne inférieur obtenu après optimisation. En rouge, nous avons tracé la courbe polynomiale obtenue et en jaune les points de contrôle de la courbe.

L'étape suivante est la recherche des commissures internes de la bouche. La méthode de recherche que nous avons employée est également analogue à celle développée pour la recherche des commissures externes. On fait l'hypothèse que, les commissures internes se trouvent sur les lignes de minimum de luminance qui sont initialisées aux positions des commissures externes (figure 5.26). La procédure de recherche des commissures internes est ensuite identique au cas des commissures externes.

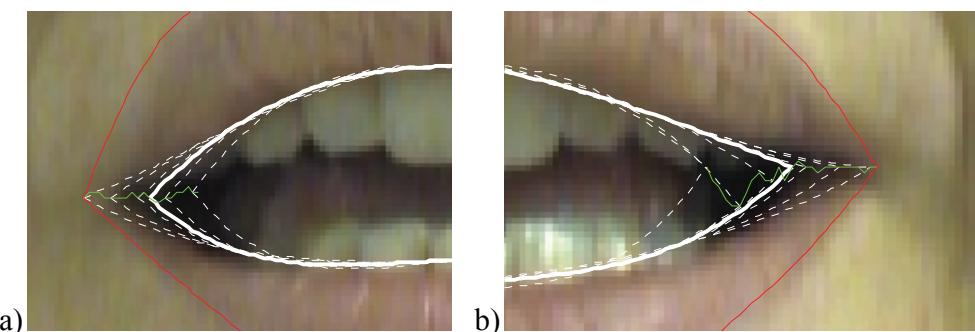


Figure 5.26 : Exemple de recherche des commissures internes pour une bouche ouverte, a) Recherche de la commissure interne gauche, b) recherche de la commissure interne droite

Enfin, à partir des positions des commissures internes et des jeux de points de contrôle optimaux des contours internes, supérieur et inférieur, on calcule les contours internes supérieur et inférieur finaux par la méthode des moindres carrés avec les contraintes que les courbes passent par les commissures internes (figure 5.27). En pratique, nous avons utilisé des ensembles de 20 points de contrôle pour estimer les contours internes.



Figure 5.27 : Exemple de contour interne final pour une bouche ouverte.

5.3.2 Modélisation du contour interne des lèvres pour le cas des bouches fermées

Dans le cas d'une bouche fermée, le problème de la modélisation du contour interne est beaucoup plus simple. Le contour interne se résume à la jointure entre les 2 lèvres. Tout comme les commissures, qui se situent dans des zones sombres de la bouche, quand la bouche est fermée, la zone qui sépare les 2 lèvres est sombre. Pour extraire le contour interne, on effectue alors un chaînage, vers la commissure droite et vers la commissure gauche, en partant du point le plus sombre, le germe, sur la colonne centrale de la bouche. On obtient la ligne L_{min} joignant les pixels sombres (figure 5.28).

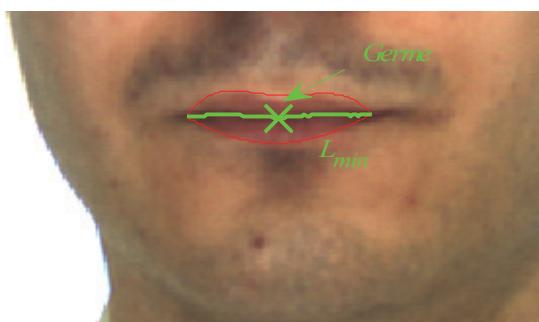


Figure 5.28 : Exemple de tracé de L_{min} pour une bouche fermée.

Pour modéliser le contour interne final d'une bouche fermée, nous avons calculé une courbe cubique par la méthode des moindres carrés à partir de L_{min} (figure 5.29).

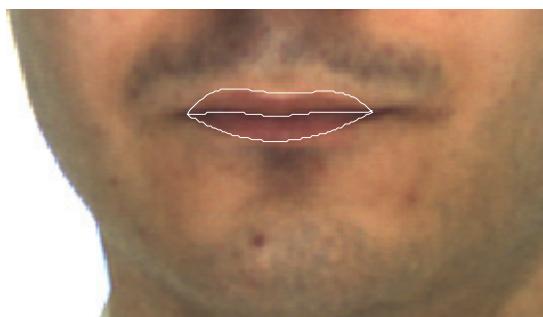


Figure 5.29 : Exemple de contour interne final pour une bouche fermée.

5.4 Résultats expérimentaux

Pour évaluer la performance de nos algorithmes de segmentation des contours de la bouche, nous avons privilégié une méthode quantitative basée sur la comparaison entre les résultats de segmentation donnés par les algorithmes et des vérités-terrain. Nous avons constitué une base de 957 images qui ont été manuellement annotées par des experts différents pour extraire le contour externe et le contour interne de la bouche. Notre base est constituée d'images présentant une grande variété de formes de bouche. De plus, nous avons veillé à ce que les images sélectionnées proviennent de sujets et de sources multiples. La base est composée des séquences suivantes :

- 6 séquences de 25 images, de 6 sujets différents, acquises à l'aide d'un casque porté par le sujet et sur lequel était fixée une micro caméra centrée sur la bouche. Chaque séquence représente la prononciation d'un phonème particulier. Sur la première ligne de la figure 5.30, on donne des exemples d'images provenant de ces séquences. On nomme cette série d'images « série 1 ».
- 100 images proviennent de 4 séquences dynamiques d'un même locuteur filmé de face par un dispositif analogue à celui utilisé pour les séquences précédentes et prononçant des numéros de téléphones. La deuxième ligne de la figure 5.30 présente des exemples tirés de ces séquences. On nomme cette série d'images « série 2 ».

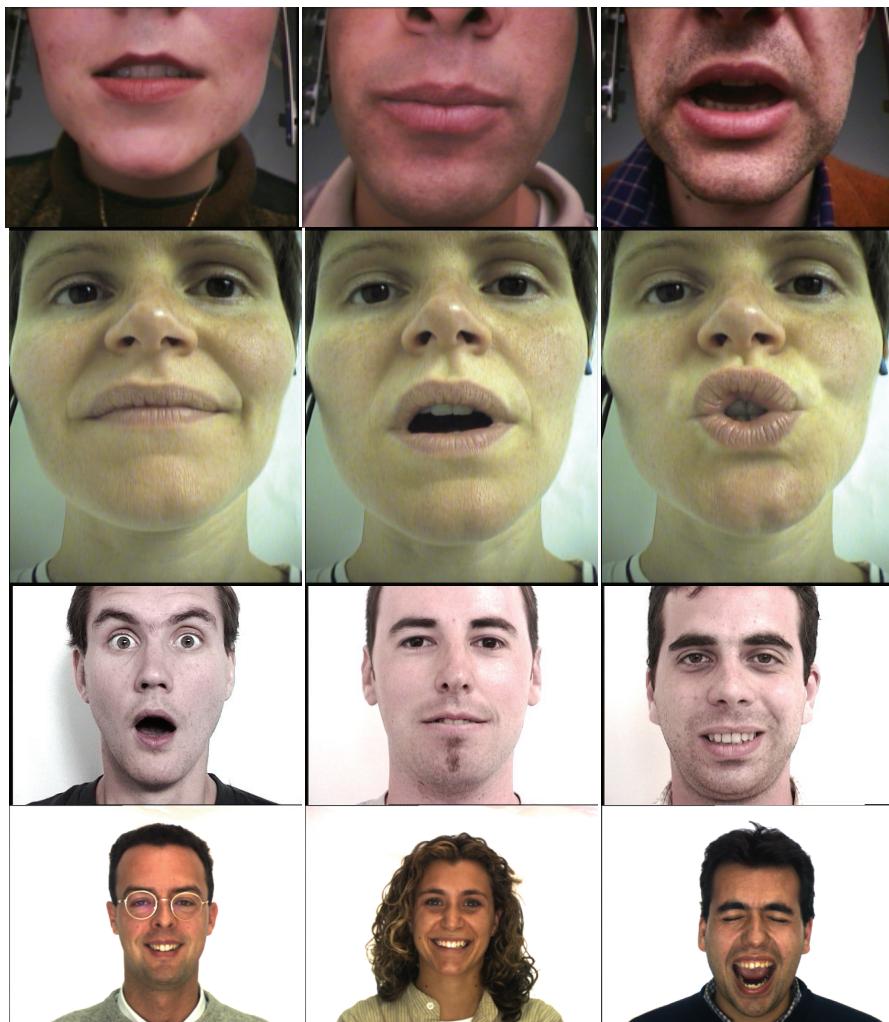


Figure 5.30 : Exemples d'images tirées des bases d'images.

- 8 séquences comprenant 25 images de visage de 5 sujets ont été filmées par une webcam. Sur ces séquences, il a été demandé au sujet de simuler une expression faciale relative à la surprise, la peur, le dégoût ou la joie en partant d'une expression neutre. La 3^{ième} ligne de la figure 5.30 montre des exemples d'images tirés de ces séquences. On nomme cette série d'images « série 3 ».
- 507 images de visage de 125 sujets différents ont été extraites de la base AR (Martinez 1998). Ces images correspondent aux formes de bouche « sourire » et « cri ». La 4^{ième} ligne de la figure 5.30 donne des exemples d'images provenant de la base AR. On nomme cette série d'images « série 4 ».

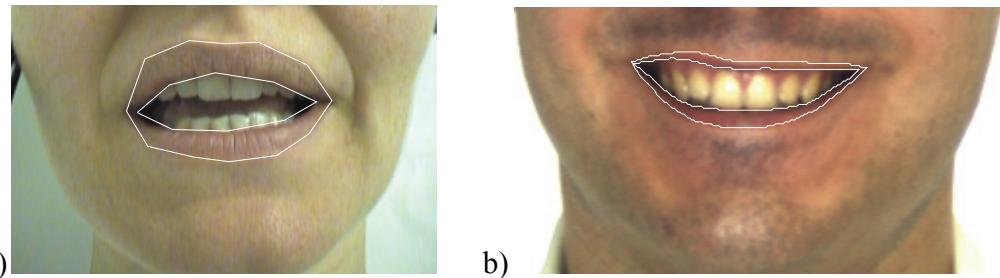


Figure 5.31 : Exemples de vérités-terrain.

Pour les séries 1, 2 et 3, les contours externes et internes ont été annotés manuellement (12 points pour le contour externe, 8 points pour le contour interne) (figure 5.31-a). Pour la série 4, la vérité-terrain a été obtenue par une méthode semi-automatique : un expert a placé manuellement des points de contrôle sur les contours externes et internes. Une optimisation a par la suite été réalisée pour obtenir les contours de référence de la bouche (figure 5.31-b).

5.4.1 Evaluation des performances basée sur le calcul d'une aire relative

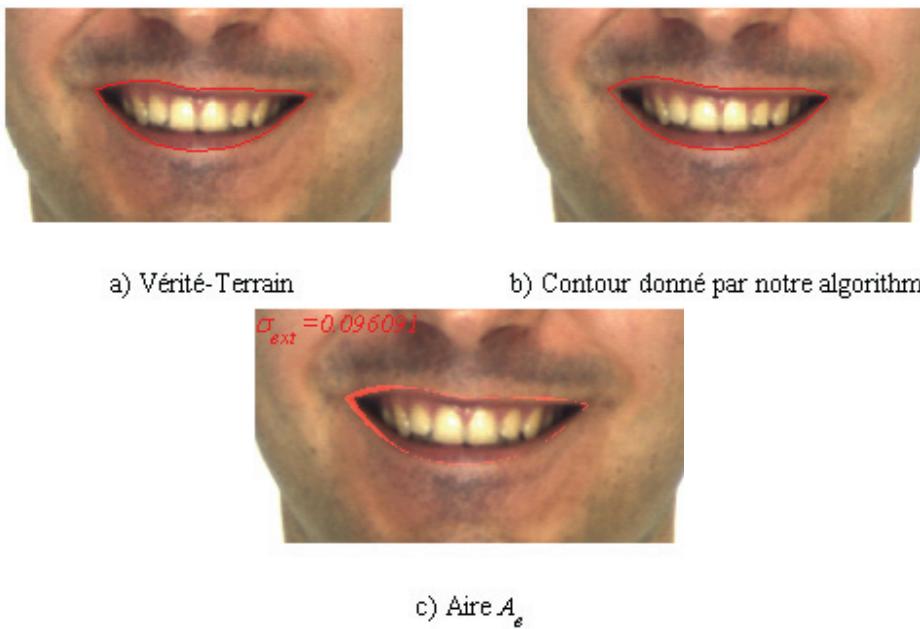


Figure 5.32 : Exemple de tracé de l'aire A_e entre le contour externe d'une bouche provenant de la vérité-terrain et le contour externe donné par nos algorithmes, on donne également σ_{ext} .

Pour comparer les contours extraits par nos algorithmes aux vérités-terrain, nous avons utilisé plusieurs critères de performances. En premier lieu, nous avons utilisé un critère basé sur l'aire A_e (figure 5.32-c), en nombre de pixels, entre les contours donnés par nos algorithmes et les contours des vérités-terrain. Cette aire est ensuite normalisée par l'aire totale définie par le contour externe ou interne de la vérité-terrain en nombre de pixels. On aboutit aux critères σ_{ext} et σ_{int} , respectivement, pour le contour externe et le contour interne de la bouche. Ces critères représentent l'erreur relative, en pourcentage, entre les contours donnés par nos algorithmes de segmentation et les vérités-terrain.

5.4.1.1 Evaluation quantitative des performances pour la segmentation du contour externe et du contour interne par les critères σ_{ext} et σ_{int}

Nous avons d'abord testé la performance de la segmentation du contour externe de la bouche. La première ligne de la table 5.2 donne la moyenne $\bar{\sigma}_{ext}$ ainsi que l'écart type du critère σ_{ext} pour les 450 images des séries 1,2 et 3 dont les contours externes ont été segmentés manuellement. La seconde ligne de la table 5.2 donne la moyenne $\bar{\sigma}_{ext}$ du critère pour les images de la série 4. Nous donnons les résultats séparément pour les séries 1, 2, 3 et 4 car les vérités-terrain ont été créés dans des conditions différentes.

$\bar{\sigma}_{ext}$ pour les séries d'images 1, 2 et 3	$10,2 \pm 5 \%$
$\bar{\sigma}_{ext}$ pour la série d'images 4	$8,5 \pm 6,5 \%$

Table 5.2 : Evaluation de la performance de la segmentation du contour externe par le critère σ_{ext} .

Les résultats de la table 5.2 montrent que les erreurs relatives moyennes pour les 2 bases d'images sont comparables. L'erreur moyenne est légèrement plus grande sur les séries 1, 2 et 3. On peut attribuer cette légère différence au fait que le contour externe a été segmenté moins précisément que pour la série 4. Sur la figure 5.32-c, nous avons également affiché la valeur σ_{ext} pour l'exemple.

Nous avons ensuite effectué l'évaluation de la performance de la segmentation du contour interne des lèvres par le critère σ_{int} . Les résultats sont donnés à la table 5.3.

$\bar{\sigma}_{int}$ pour les séries d'images 1, 2 et 3	$23 \pm 9 \%$
$\bar{\sigma}_{int}$ pour la série d'images 4	$17.2 \pm 12.5 \%$

Table 5.3 : Evaluation de la performance de la segmentation du contour interne par le critère σ_{int} .

On constate une augmentation de l'erreur relative dans le cas de la segmentation du contour interne de la bouche. Ce comportement illustre le problème rencontré lorsqu'on effectue une évaluation par le calcul d'une erreur relative. La référence qui est utilisée pour l'évaluation provient d'une segmentation manuelle, donc subjective par définition. Cette segmentation donne une idée assez précise du contour recherché, mais pas une référence absolue. Une vérité-terrain produite par un expert différent pourra donner des résultats différents. Ce type d'erreur relative ne nous indique pas si les formes extraites sont visuellement proches des vérités-terrain. Pour le cas du contour interne, les ouvertures de la bouche peuvent être faibles, et une erreur absolue (en pixels) faible peut engendrer une erreur relative très importante. La figure 5.33 illustre ce phénomène. L'erreur relative σ_{int} est de 30 % alors que visuellement la segmentation automatique semble correcte.



Figure 5.33 : Tracés des contours internes donnés par la vérité-terrain et par notre algorithme. Le tracé rouge correspond à la segmentation automatique et le tracé blanc correspond à la vérité-terrain. On donne également σ_{int} .

La figure 5.33 montre bien que ce critère basé sur une erreur relative ne permet pas de quantifier la ressemblance entre la vérité-terrain et le contour segmenté automatiquement. Inversement, pour une bouche grande ouverte l'erreur relative peut être faible alors que, visuellement le contour est mal segmenté. Sur la figure 5.34, nous présentons une segmentation erronée du contour interne de la bouche. Visuellement l'erreur semble plus importante que pour le cas de la figure 5.33, pourtant l'erreur relative est inférieure ($\sigma_{int}=17.2 \%$).



Figure 5.34 : Tracés des contours internes donnés par la vérité-terrain et par notre algorithme. Le tracé rouge correspond à la segmentation automatique et le tracé blanc correspond à la vérité-terrain. On donne également σ_{int} .

5.4.2 Calcul des descripteurs de Fourier pour l'évaluation de la performance de la segmentation

Pour évaluer de manière plus pertinente la performance de nos algorithmes de segmentation, nous nous sommes intéressés aux descripteurs de Fourier. Les descripteurs de Fourier permettent de décrire la forme d'un objet tout en étant invariants aux rotations, aux translations et aux changements d'échelles (Zahn, 1972; Granlund, 1972). Dans (Granlund, 1972) l'auteur a introduit une méthode basée sur une représentation complexe pour calculer les descripteurs de Fourier d'un contour caractérisé par N_p points. Le contour est représenté par l'ensemble $Z=\{z_i=x_i+jy_i, i=1,\dots,N_p\}$ avec N_p points et $\{x_i, y_i\}$ les coordonnées des points échantillonnant le contour. Les descripteurs de Fourier $\{C_k, k=-N_p/2+1, \dots, N_p/2\}$ sont les coefficients complexes de la transformée de Fourier de Z .

$$z_i = \sum_{k=-N_p/2+1}^{N_p/2} C_k \exp\left(2\pi j \frac{ki}{N_p}\right) \quad (53)$$

La relation inverse liant les coefficients C_k aux z_i est alors de la forme suivante :

$$C_k = \frac{1}{N_p} \sum_{i=1}^{N_p} z_i \exp\left(-2\pi j \frac{ki}{N_p}\right) \quad (54)$$

La restriction des C_k à $k=-N_p/2+1, \dots, N_p/2$ vient du fait que la fréquence maximale, d'après le théorème de Shannon, est obtenue pour $k=N_p/2$. Les descripteurs de Fourier vont donc décrire le spectre des fréquences d'un contour donné. Les coefficients C_k avec k proche de zéro décriront les basses fréquences. Ces coefficients nous renseigneront sur la forme approximative du contour. Les C_k , avec k élevé, nous renseigneront sur les hautes fréquences du contour, donc sur les détails de la forme :

- Pour $k=0$, d'après (53), le coefficient C_0 correspond au centre de gravité du contour. Ce terme n'affecte pas la description de la forme du contour. En retirant ce terme de l'ensemble des descripteurs, on obtient une description du contour invariante aux translations.
- Le coefficient C_1 décrit l'échelle du contour. Plus précisément, si tous les coefficients C_k , excepté C_1 , sont fixés à 0, et si l'on effectue la transformée inverse pour retrouver le contour correspondant, alors on obtient un cercle (en fait un polygone) composé de N_p points. Le module de C_1 nous donne le rayon du cercle. En normalisant les autres coefficients par le module de C_1 , on obtient alors une description de la forme du contour indépendante de l'échelle.
- Les autres coefficients, $k \neq \{0,1\}$, vont définir les altérations du cercle défini par C_1 . Les coefficients C_k , avec $k > 1$, auront pour effet de déformer le cercle en « poussant » vers l'extérieur les points avec une périodicité de $k-1$. Pour les C_k avec $k < 0$, l'effet est inverse : le cercle caractérisé par le module de C_1 sera déformé vers l'intérieur en « tirant » les points avec une périodicité de $1-k$. La phase des C_k va quant à elle nous renseigner sur la localisation des déformations sur le cercle de base. En calculant le spectre en amplitude des coefficients C_k , nous obtiendrons une description des déformations du contour.

Notre objectif dans cette section est de proposer une méthode d'évaluation de nos algorithmes de segmentation, par rapport à une vérité-terrain, plus pertinente que le calcul d'une erreur relative basée sur des aires. Pour évaluer la performance de notre segmentation, nous proposons de comparer les descripteurs de Fourier des contours donnés par les vérités-terrain et nos algorithmes de segmentation. La première étape est de ré-

échantillonner les contours externes et internes donnés par les vérités-terrain et par nos algorithmes de segmentation. Le but est que tous les contours, des vérités-terrain et ceux segmentés par nos algorithmes, aient le même nombre de points N_p . Pour les contours externes, le nombre N_{pext} a été fixé au nombre de points moyen des contours externes calculé sur l'ensemble de nos images de bouche (80 points). De la même manière, nous avons calculé N_{pint} pour ré-échantillonner les contours internes (50 points). Pour évaluer la performance de nos algorithmes, nous proposons le calcul des critères suivants pour les 2 catégories de contours (externe et interne) :

- Calcul de l'erreur moyenne, en pixels, entre les centres de gravité des contours donnés par les vérités-terrains Cvt_0 et par la segmentation Cs_0 .
- Calcul de l'erreur moyenne, en pixels, entre les rayons des cercles donnés par les modules des descripteurs Cvt_1 et Cs_1 , respectivement, des vérités-terrain et des résultats des segmentations.
- Ensuite pour tous les contours, on normalise les ensembles de descripteurs pour $\{k=-N_p/2+1, \dots, -1, 2, \dots, N_p/2\}$ par le module du descripteur C_1 correspondant afin de former des spectres normalisés en amplitude pour les vérités-terrain et les contours segmentés par nos algorithmes (figure 5.35). On calcule ensuite la corrélation (comprise entre 0 et 1) entre les spectres d'amplitude normalisés des contours des vérités-terrain et des contours donnés par nos algorithmes de segmentation pour chaque image. Enfin, on calcule la corrélation moyenne sur l'ensemble des images.

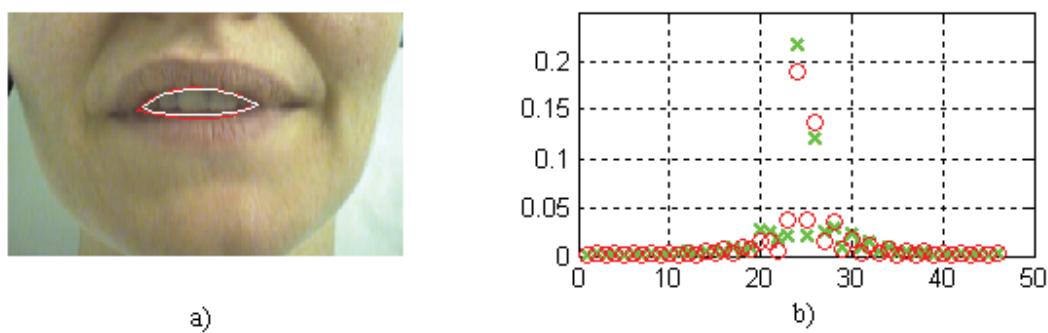


Figure 5.35 : Exemple de tracé de spectres d'amplitude normalisés des descripteurs de Fourier, pour $\{k=-N_p/2+1, \dots, -1, 2, \dots, N_p/2\}$, du contour interne de la vérité terrain et du contour obtenu par nos algorithmes de segmentation, a) Tracés de la vérité-terrain (blanc) et du résultat de la segmentation (rouge), b) Tracés des spectres d'amplitude normalisés pour la vérité-terrain (croix vertes) et pour le contour obtenu par segmentation (rond rouge).

Les résultats pour la segmentation du contour externe de la bouche pour les séries 1, 2, 3 sont donnés à la table 5.4. Les résultats de la comparaison pour la série 4 sont donnés à la table 5.5. Dans les tables qui suivent, nous avons également inclus la surface moyenne entourée par les contours externes des vérités-terrain afin de mettre en perspective les 2 premiers critères qui sont des erreurs absolues exprimées en pixels.

Distance moyenne entre Cvt_0 et Cs_0 (en pixels).	1.7 ± 1
Erreur moyenne entre Cvt_1 et Cs_1 (en pixels).	0.83 ± 1.25
Corrélation moyenne entre les spectres d'amplitude normalisés des descripteurs de Fourier donnés par les contours des vérités-terrain et par les contours obtenus par nos algorithmes de segmentation.	0.97 ± 0.025
Surface moyenne entourée par les contours externes donnés par les vérités-terrain (en pixels)	3296 ± 2102

Table 5.4 : Evaluation de la performance de la segmentation du contour externe par comparaison des descripteurs de Fourier pour les images des séries 1, 2, 3.

Distance moyenne entre Cvt_0 et Cs_0 (en pixels).	2.7 ± 2.1
Erreur moyenne entre Cvt_1 et Cs_1 (en pixels).	1.1 ± 1.2
Corrélation moyenne entre les spectres d'amplitude normalisés des descripteurs de Fourier donnés par les contours des vérités-terrain et par les contours obtenus par nos algorithmes de segmentation.	0.98 ± 0.026
Surface moyenne entourée par les contours internes donnés par les vérités-terrain (en pixels)	5580 ± 2580

Table 5.5 : Evaluation de la performance de la segmentation du contour externe par comparaison des descripteurs de Fourier pour les images de la série 4.

Pour le cas de la segmentation du contour interne de la bouche, nous donnons les résultats des comparaisons entre les vérités-terrain et les contours donnés par nos algorithmes aux tables 5.6 et 5.7, respectivement, pour les séries 1, 2, 3 et pour la série 4.

Distance moyenne entre Cvt_0 et Cs_0 (en pixels).	1.7 ± 2
Erreur moyenne entre Cvt_1 et Cs_1 (en pixels).	1.1 ± 1.6
Corrélation moyenne entre les spectres d'amplitude normalisés des descripteurs de Fourier donnés par les contours des vérités-terrain et par les contours obtenus par nos algorithmes de segmentation.	0.97 ± 0.025
Surface moyenne entourée par les contours externes donnés par les vérités-terrain (en pixels)	770 ± 718

Table 5.6 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour les images des séries 1, 2, 3.

Distance moyenne entre Cvt_0 et Cs_0 (en pixels).	2.8 ± 2.53
Erreur moyenne entre Cvt_1 et Cs_1 (en pixels).	2.9 ± 2.3
Corrélation moyenne entre les spectres d'amplitude normalisés des descripteurs de Fourier donnés par les contours des vérités-terrain et par les contours obtenus par nos algorithmes de segmentation.	0.97 ± 0.036
Surface moyenne entourée par les contours internes donnés par les vérités-terrain (en pixels)	3263 ± 2101

Table 5.7 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour les images de la série 4.

Pour les cas particuliers des contours internes des figures 5.33 et 5.34, les résultats sont donnés, respectivement, par les tables 5.8 et 5.9.

Distance entre Cvt_0 et Cs_0 (en pixels).	2.1
Erreur entre Cvt_1 et Cs_1 (en pixels).	2.14
Corrélation entre le spectre d'amplitude normalisé calculé à partir du contour donné par la vérité-terrain et le spectre normalisé calculé à partir du résultat de la segmentation.	0.99
Surface entourée par le contour interne donné par la vérité-terrain (en pixels)	1896

Table 5.8 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour l'image de la figure 5.33.

Distance entre Cvt_0 et Cs_0 (en pixels).	3.37
Erreur entre Cvt_1 et Cs_1 (en pixels).	4.6791
Corrélation entre le spectre d'amplitude normalisé calculé à partir du contour donné par la vérité-terrain et le spectre normalisé calculé à partir du résultat de la segmentation.	0.78
Surface entourée par le contour interne donné par la vérité-terrain (en pixels)	5707

Table 5.9 : Evaluation de la performance de la segmentation du contour interne par comparaison des descripteurs de Fourier pour l'image de la figure 5.34.

Pour les cas particuliers des figure 5.33 et 5.34, on remarque que le critère basé sur la corrélation des spectres d'amplitudes des C_k pour $\{k=-N_p/2+1, \dots, -1, 2, \dots, N_p/2\}$ est bien plus faible pour l'image de la figure 5.34 que pour l'image de la figure 5.33. De même, les erreurs entre Cvt_0 et Cs_0 et entre Cvt_1 et Cs_1 sont plus importantes pour le cas de la figure 5.34. La comparaison indique clairement que la segmentation du contour interne est moins précise sur la figure 5.34. La comparaison des contours par les descripteurs de Fourier est plus pertinente que le simple calcul d'une erreur relative.

Les figures 5.36, 5.37, 5.38, 5.39 et 5.40 présentent des exemples de segmentation du contour externe et interne sur des bouches ouvertes. Ces exemples permettent d'illustrer la flexibilité de nos algorithmes de segmentation.



Figure 5.36 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 1.

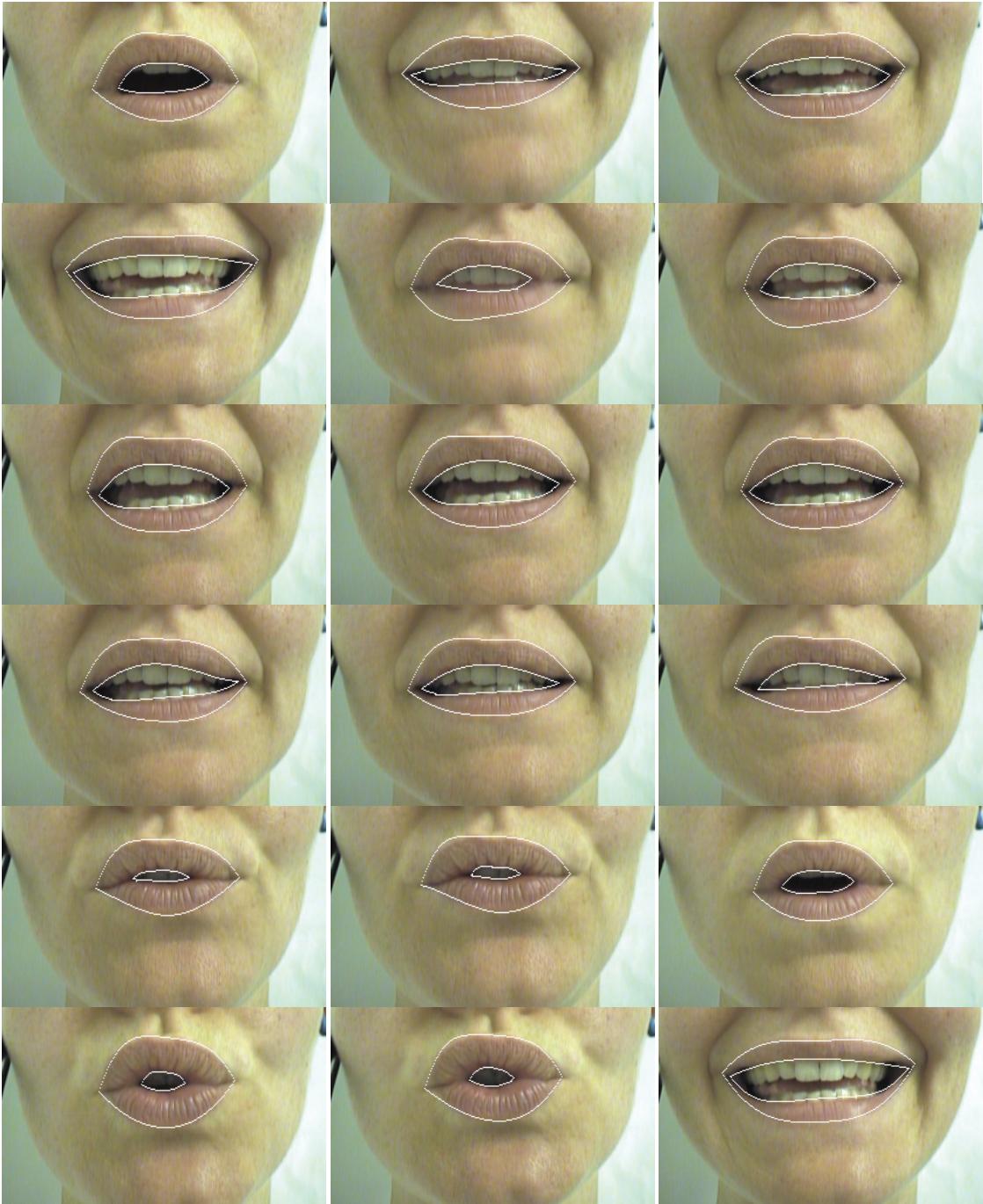


Figure 5.37 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 2.



Figure 5.38 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 3.

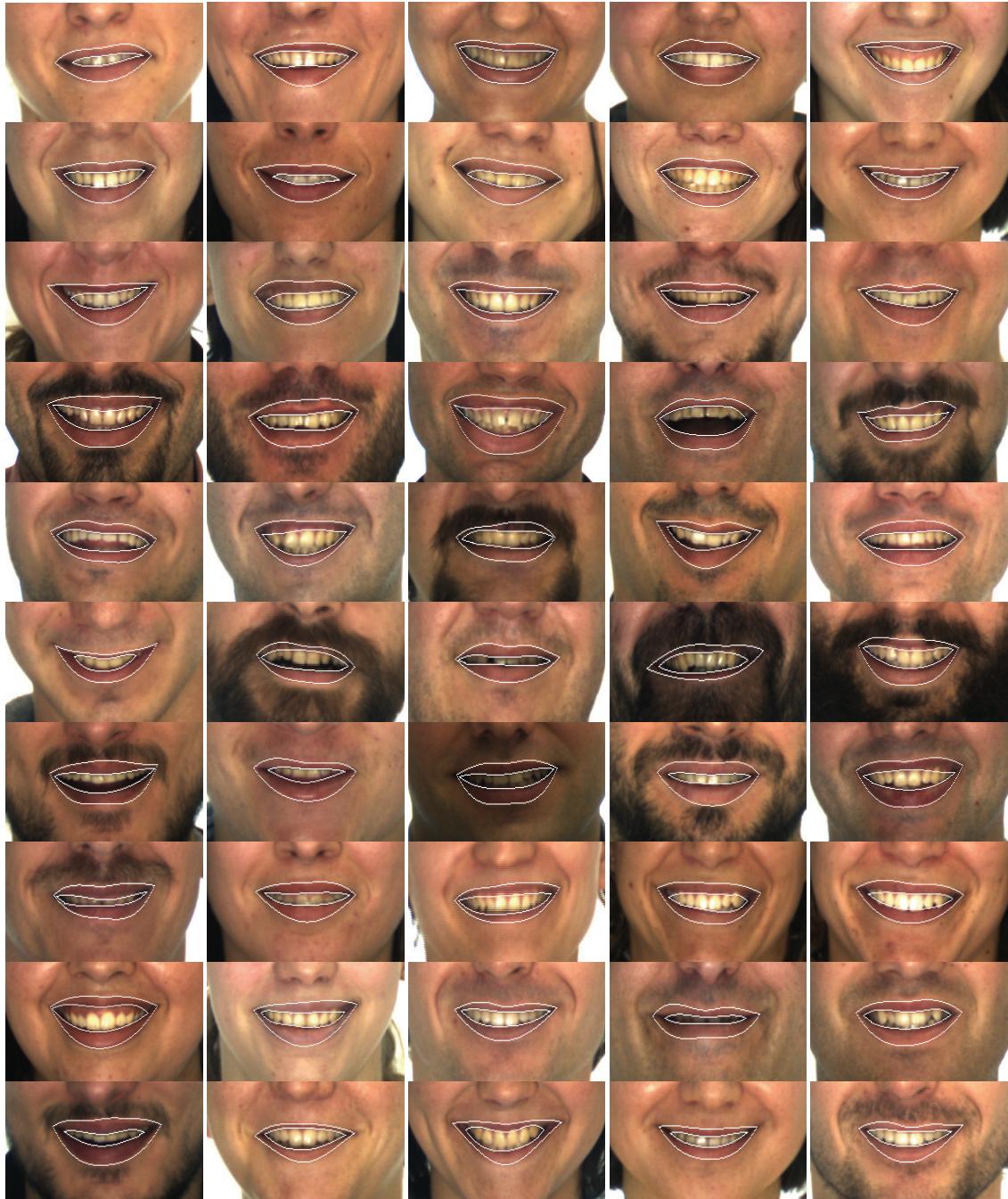


Figure 5.39 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 4.



Figure 5.40 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de la série 4 avec des ouvertures extrêmes de la bouche.

La figure 5.41 présente des exemples de segmentation du contour externe et du contour interne pour des images de bouche fermée.



Figure 5.41 : Exemples de segmentation du contour externe et du contour interne des lèvres pour des images de bouche fermée.

5.4.3 Cas limites

Nous avons proposé dans la section précédente une méthode d'évaluation de la performance de nos algorithmes de segmentation ainsi que les résultats obtenus lors de nos simulations. Nous nous étions fixés comme objectif de produire une segmentation robuste et précise du contour externe et du contour interne des lèvres. Les résultats de la section 5.4.2 montrent que nos algorithmes de segmentation des contours des lèvres sont robustes et offrent une bonne précision. Toutefois, durant nos simulations, nous avons observé que des erreurs de segmentation pouvaient se produire dans certains cas.

En ce qui concerne le contour externe de la bouche, la segmentation est robuste même en présence de barbe et de moustache et même lorsque les lèvres sont partiellement occultées (voir les figures 5.36, 5.37, 5.38, 5.39, 5.40 et 5.41). Des erreurs de segmentation sont toutefois apparues dans quelques cas où une des 2 lèvres était presque totalement occultée (figure 5.42-a).

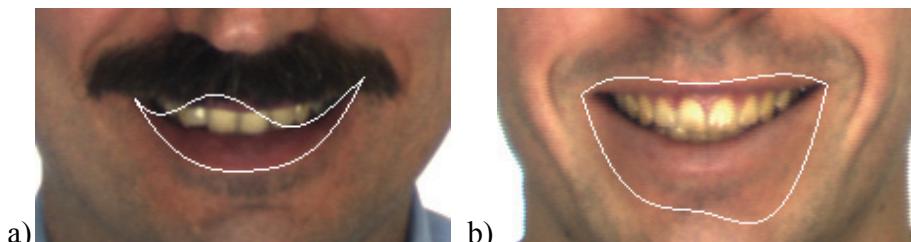


Figure 5.42 : Exemples de segmentations erronées du contour externe.

La segmentation peut également échouer dans des cas où les contours des lèvres sont presque invisibles (figure 5.42-b). Dans l'exemple de la figure 5.42-b, le contour externe de la lèvre inférieure est très peu marqué et l'intensité des gradients, même pour plusieurs échelles, n'est pas assez grande pour que le contour s'optimise correctement. Toujours en ce qui concerne la segmentation du contour externe, nous avons également testé nos algorithmes sur des images de sujets ayant la peau noire. En effet pour notre étude colorimétrique sur le contraste peau/lèvres, les bases d'images dont nous disposions incluaient essentiellement des individus de type indo-européens. Nous avons donc testé nos

algorithmes sur une trentaine d'images extraites de la base FERET (Philipps, 2000 ; Feret). Des exemples sont présentés à la figure 5.43.



Figure 5.43 : Exemples de segmentation correcte avec des sujets ayant la peau noire.

Des problèmes de segmentation du contour externe de la bouche se sont produits avec des sujets pour lesquels il n'y avait pas de différence de teinte entre la peau et les lèvres malgré le traitement par l'algorithme allongement-décorrélation. La figure 5.44 illustre le cas de figure que nous avons rencontré avec des images extraites de la base FERET. On constate que la teinte de la lèvre supérieure est identique à celle de la peau. Dans ces conditions, nous allons segmenter la lèvre inférieure.

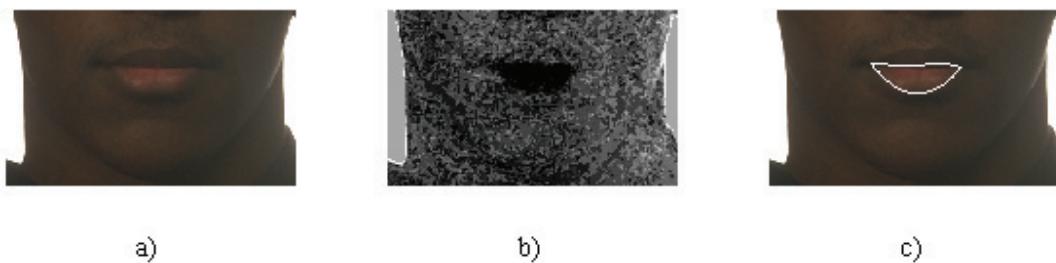


Figure 5.44 : Exemple de segmentation erronée avec un sujet ayant la peau noire, a) image d'entrée, b) teinte \hat{U} , c) Contour externe segmenté.

Ce type de cas de figure peut également se produire avec des sujets de couleur de peau blanche comme à la figure 5.42-b. Un axe de recherche serait donc de constituer une base plus conséquente de sujets de couleur de peau noire et de répéter l'étude colorimétrique pour essayer de dégager une grandeur colorimétrique pertinente, ou plus spécifiquement, une méthode pour traiter ces sujets. Par manque de temps, nous n'avons pas pu entreprendre ces démarches.

Ensuite, du point de vue de la segmentation du contour interne, il y a également quelques cas de figure qui ont provoqué des segmentations erronées. Il faut noter que cette étape est plus délicate que le cas de la segmentation du contour externe à cause du grand nombre de configurations possibles que peut prendre la zone interne de la bouche. Le nombre de contours parasites est potentiellement plus important. Pour des bouches présentant une ouverture moyenne, comme celles des figures 5.36, 5.37, 5.38 et 5.39 les erreurs sont survenues dans certains cas où les gencives sont visibles et fines. Le contraste lèvre/gencive est très faible. Le contour interne supérieur aura tendance à converger sur la transition gencive/dents. Les dents présentent un fort contraste avec les gencives (figure 5.45).



Figure 5.45 exemples de segmentations erronées du contour interne supérieur dues à la présence des gencives.

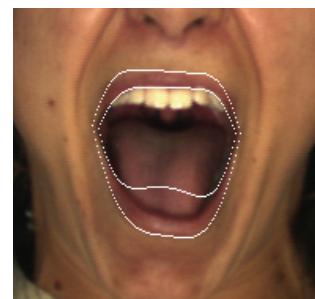


Figure 5.46 : Exemple de segmentation erronée du contour interne inférieure.

Enfin, pour les bouches présentant des ouvertures extrêmes, lorsque la langue est visible, étant donné que sa teinte est très proche de celle des lèvres, le contour peut être attiré vers l'intérieur de la bouche, en fait vers les zones les plus sombres, comme sur la figure 5.46.

5.5 Bilan

Dans ce chapitre, nous avons proposé un ensemble d'algorithmes pour modéliser le contour externe ainsi que le contour interne des lèvres. Les algorithmes proposés reposent sur une localisation préalable de la bouche, en particulier des lèvres, ainsi que sur les gradients γ -normalisés, dans l'optique de maximiser la robustesse aux variations d'échelle et d'éclairage. Notre but était également de nous affranchir des étapes fastidieuses de réglage des paramètres qui interviennent lors de l'initialisation des snakes, tout en gardant une grande flexibilité au niveau des formes. Pour cela, nous avons proposé une méthode hiérarchique utilisant des polynômes pour modéliser les contours des lèvres. L'intérêt de cette méthode réside dans l'adaptation automatique de la complexité des courbes qui permet de maximiser la précision de la segmentation.

Pour évaluer la performance de la segmentation des contours de la bouche, nous avons proposé une méthode basée sur les descripteurs de Fourier. Cette approche permet d'effectuer une comparaison plus pertinente qu'un calcul d'erreur relative entre les résultats des segmentations automatiques et des vérités-terrain. Cette évaluation montre que les objectifs de robustesse et de précision que nous nous sommes fixés au début de nos travaux sont atteints.

Du point de vue de la complexité, pour une image de bouche d'une résolution de 102x170 pixels, l'optimisation du contour externe a pris 15s et 10s pour le contour interne sur une machine équipée d'un processeur double cœur avec une fréquence d'horloge de 2.33GHz et accompagné de 2 Go de mémoire RAM. Il faut préciser que nos algorithmes ont été implémentés sous Matlab.

Conclusion et perspectives

Dans ce manuscrit, nous avons présenté l'ensemble de nos travaux de thèse sur la modélisation de la zone de la bouche. Cette thèse a été motivée par les recherches effectuées au GIPSA-lab depuis le début des années 90 sur la modélisation de la bouche. Nous avons pu voir que la modélisation de la bouche était requise dans de nombreuses applications telles que les projets TEMPOVALSE (Bailly, 2003), TELMA (Beautemps, 2007) ou encore le système de maquillage automatique de la société Vesalis.

Les objectifs visés étaient de proposer un ensemble de méthodes permettant de modéliser précisément la zone de la bouche sur des images statiques, avec la meilleure robustesse et la meilleure précision possible. Par robustesse, nous entendions obtenir une méthode fiable ne nécessitant pas de réglage de paramètres. Par précision, nous entendions fournir une modélisation fidèle des contours de la bouche.

Travail réalisé

Au travers de ce rapport, nous avons présenté notre méthode de segmentation «région-contour» pour extraire les contours externes et internes des lèvres. Au chapitre 2, nous avons présenté la première étape de notre méthode de modélisation de la bouche. Cette étape consiste en une localisation des lèvres sur des images de visage à l'aide d'une teinte calculée pour maximiser le contraste peau/lèvres. Nous avons mené une étude sur les grandeurs colorimétriques adaptées à l'étude des lèvres. Après avoir proposé l'utilisation de l'algorithme allongement-décorrélation pour augmenter le contraste peau/lèvres, notre étude a permis de montrer l'intérêt de la grandeur colorimétrique \hat{U}_{decorr} pour séparer la peau et les lèvres. Par la suite, nous avons montré l'intérêt d'un formalisme multi-échelles, inspiré des travaux de (Lindeberg 1998), pour caractériser les contours des lèvres. Enfin, une méthode de seuillage automatique exploitant les gradients γ -normalisés et la teinte \hat{U}_{decorr} a été introduite pour segmenter les lèvres.

Au chapitre 3, un nouveau système supervisé de détection de l'état ouvert/fermé de la bouche a été présenté. Le but était de permettre un traitement distinct des contours internes pour les bouches ouvertes et fermées. Pour réaliser la détection de l'état de la bouche, une approche fréquentielle, basée sur un modèle du système visuel humain, a été privilégiée. L'intérêt de ce type d'approche réside dans le fait que nous avons cherché à modéliser une chaîne de traitement (dans notre cas la rétine et le cortex V1) dont on sait qu'elle est

cohérente et performante. Cette approche nous a permis d'obtenir une détection robuste et rapide de l'état de la bouche.

Toujours dans le but d'améliorer la robustesse de la segmentation de la bouche, au chapitre 4, nous avons étudié l'intérêt de la modalité infrarouge dans le cadre de la cotutelle de thèse avec l'Université Laval. Cette étude nous a permis de nous familiariser avec le domaine de la thermographie infrarouge et avec les contraintes liées à ces techniques d'imagerie. Les résultats de cette étude nous ont amené à faire des compromis entre les propriétés en émission du visage et les contraintes imposées par les systèmes d'acquisition pour produire une base d'images de visage combinée visible/infrarouge. Si l'étude sur le cas de la séparation peau/lèvres dans la modalité infrarouge n'a pas été concluante, elle nous a néanmoins conduits à la construction d'une base d'images qui pourra s'avérer intéressante pour des travaux d'analyse faciale nécessitant les 2 modalités.

Enfin, au chapitre 5, nous avons introduit notre méthode de segmentation hiérarchique de contours basée sur des polynômes et sur les gradients γ -normalisés pour extraire les contours des lèvres. L'intérêt de cette méthode vient du fait qu'elle ne nécessite pas de réglages préalables : la complexité des polynômes chargés de modéliser les contours est adaptée itérativement afin d'offrir une modélisation des contours des lèvres la plus fidèle possible. Nous avons proposé une évaluation quantitative de nos algorithmes de segmentation des contours des lèvres en utilisant des vérités-terrain comme référence. Une première évaluation basée sur des calculs d'erreurs relatives entre les contours segmentés et les contours des vérités-terrain a été conduite. Lors de nos simulations, les limitations de cette méthode d'évaluation nous sont apparues rapidement. Les erreurs relatives calculées ne reflétaient pas la proximité visuelle qu'il pouvait y avoir entre les contours segmentés et ceux des vérités-terrain (cf. les figures 5.33 et 5.34). Une méthode reposant sur le calcul des descripteurs de Fourier a été développée pour comparer les contours issus de nos algorithmes de segmentation et les contours issus des vérités-terrain. Le calcul de 3 critères a été proposé à partir des descripteurs de Fourier. Ces critères se sont révélés plus pertinents pour comparer les contours des lèvres que les calculs simples d'erreurs relatives. Ils montrent que nos algorithmes de segmentation offrent une bonne précision et une bonne robustesse. Enfin, une analyse des cas limites a été proposée.

Perspectives

L'analyse des cas limites nous a permis de voir que la segmentation pouvait échouer, en particulier pour des personnes de couleur de peau noire. Concernant les sujets de couleur de peau noire, le peu d'images disponibles n'a pas permis de mener une étude complète. La constitution d'une base de données apparaît donc comme indispensable pour poursuivre les travaux sur l'analyse labiale. Un travail sur notre code est également à envisager pour accélérer la vitesse de traitement. Actuellement nos algorithmes sont implantés sous Matlab et sont à l'état de prototype. Une étape de « nettoyage » du code paraît indispensable pour optimiser le code avant d'envisager une implémentation sous un autre langage.

Toujours en ce qui concerne les axes de recherche possibles, en nous inspirant de la méthode que nous avons développée au chapitre 3 pour la détection de l'état ouvert/fermé de la bouche, nous pensons qu'il serait possible de détecter directement des formes de bouches sans passer par la segmentation des contours. Au cours de nos travaux, nous avons essayé de combiner les informations données par les spectres log-polaires et les contours segmentés manuellement, sans obtenir de résultats intéressants. Nous avons toutefois remarqué un phénomène intéressant. Lorsqu'on exécute un algorithme des K-moyennes avec un grand nombre de classes sur les spectres Log-polaires des images de bouche, les bouches qui visuellement se ressemblent (présence de dents, de la langue, sourire, ...), se regroupent dans les mêmes classes. Il est intéressant de noter que ces appariements ont lieu pour des sujets ayant des morphologies différentes. Par ailleurs, nous avons vu lors de l'évaluation de la performance qu'une comparaison directe entre les vérités-terrain et les résultats de segmentation n'était pas pertinente et que l'utilisation des descripteurs de Fourier permettait une meilleure comparaison. Plus particulièrement, l'ensemble des modules des descripteurs C_k , avec $k \neq \{0,1\}$, normalisé par le module de C_1 , fournit une description normalisée de la forme du contour (cf. section 5.4.2), tout comme les spectres Log-polaires offrent une description normalisée de la zone de la bouche. L'idée serait alors de développer une méthode pour combiner les spectres Log-polaires et les descripteurs de Fourier des contours pour obtenir une modélisation directe de la forme de la bouche. L'intérêt de cette approche serait de lier directement l'apparence et la forme par un modèle non-linéaire (un réseau *SVM* par exemple) dans le but de s'affranchir de la classique étape

de relaxation du modèle que l'on retrouve avec la plupart des méthodes de modélisation de la bouche (*ASM*, *AAM*, snakes, modèles paramétriques).

Bibliographie

Aizerman, M., Braverman, E., Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning , Automation and Remote Control, vol. 25, p. 821-837.

Aleksic, P.S., Williams, J.J., Wu, Z. & Katsaggelos, A.K. (2002). Audiovisual speech recognition using MPEG-4 compliant visual features, EURASIP Journal on Applied Signal Processing, special Issue on Audio-Visual Speech Processing, pp. 1213-1227.

ARbase The AR face database home page

http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html

Bailly, G., Elisei, F., Odisio, M., Pele, D., Caillier, D., & Grein-Cochard, K. (2003). Talking faces for MPEG-4 compliant scalable face-to-face telecommunication, Proceedings of the Smart Objects Conference, pp. 204–207, Grenoble, France.

Ballerini, L. (1999). Genetic Snakes for Medical Images Segmentation, Lecture Notes in Computer Science, vol. 1596, pp. 59-73.

Barnard, M., Holden, E.J. & Owens, R. (2002). Lip Tracking using Pattern Matching Snakes. The 5th Asian Conference on Computer Vision (ACCV'2002), pp. 273-278, Melbourne, Australia.

Beaudot, W. (1994). Le traitement neuronal de l'information dans la rétine des vertébrés : Un creuset d'idées pour la vision artificielle. Thèse de doctorat, Institut Polytechnique de Grenoble.

Beaumesnil, B., Chaumont, M. & Luthon F. (2006). Liptracking and MPEG4 Animation with Feedback Control, IEEE International Conference On Acoustics, Speech, and Signal Processing, (ICASSP'2006), Vol. 2, pp. 677-680, Toulouse, France.

Beaumesnil, B. (2006). Suivi labial couleur pour analyse-synthèse vidéo et communication temps-réel. Thèse de doctorat, Université de Pau et des Pays de l'Adour.

Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.-A., Chne, D., Clarke, J., Elisei, F., Govokhina, O., Le, M., Marthouret, V.-B., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.-F., Tribout, M. & Vidal, S. (2007). Telma : Telephony for hearing-impaired people. From models to user tests, Proceedings of the First International Conference on Accessibility and Assitive Technology for People in Disability Situation, Toulouse, France.

Benoit, A., Caplier, A. (2005). Hypo-vigilance Analysis: Open or Closed Eye or Mouth? Blinking or Yawning Frequency ? IEEE AVSS, pp. 207-212, Como, Italie.

Benoit, A. (2007). Le système visuel humain au secours de la vision par ordinateur, Thèse de Doctorat, INPG (France).

Blake, A., Isard, M.A. & Reynard, D. (1995). Learning to track the visual motion of contours, Artificial Intelligence, pp. 101-134.

Blakemore, C, Campbell, F.W, (1969), On the existance of neurones in the human visual system selectively sensitive to the orientation and size of retinal images, Journal of Physiology, 203, 237-260.

Bouvier, C., Coulon, P.Y. & Malague, X. (2007). Unsupervised Lips Segmentation Based on ROI Optimisation and Parametric Model, International Conference on Image Processing, ICIP2007.

Bibliographie

-
- Bouvier, C., Benoit, A., Caplier A., Coulon P-Y. (2008). Open or Closed Mouth State Detection: Static Supervised Classification Based on Log-Polar Signature. Advanced Concepts for Intelligent Vision Systems, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 5259/2008, p. 1093-1102.
- Brand, J. (2001). Visual speech for speaker recognition and robust face detection. Thèse de doctorat, University of Wales, Royaume-Uni.
- Bullier, J. (2001). Integrated model of visual processing, Brain Research, vol. 36, no. 2, pp. 96-107.
- Caplier A., Sillittano S., Bouvier C., Coulon P-Y. (2009) Lip Modelling And Segmentation. Visual Speech Recognition, Visual Speech Recognition: Lip Segmentation and Mapping, Dr. Alan Wee-Chung Liew (Griffith University, Australia) and Dr. Shilin Wang (Shanghai Jiaotong University, China) (Ed.), pp. 70-127.
- Castañeda, B., Cockburn J. C. (2005). Reduced support vector machine applied to real-time face tracking, Proc. ICASSP, Vol 2, pp. 673-676, Philadelphie, États-Unis.
- Chan, M.T., Zhang, Y. & Huang, T.S. (1998). Real-time lip tracking and bimodal continuous speech recognition, Proc. IEEE Signal Processing Society Workshop on Multimedia Signal Processing, pp. 65–70, Los Angeles, États-Unis.
- Chan, M.T. (2001). HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features, Workshop on Multimedia Signal Processing, Cannes, France.
- Chen, S.C., Shao, C.L., Liang, C.K., Lin, S.W., Huang, T.H., Hsieh, M.C., Yang, C.H., Luo, C.H. & Wu, C.M.. (2004) A Text Input System Developed by using Lips Image Recognition based LabVIEW for the seriously disabled, International

Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS'2004), vol.2, pp. 4940-4943.

Chen, Q.C., Deng, G.H., Wang, X.L. & Huang, H.J. (2006) An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech, International Conference on Machine Learning and Cybernetics, pp. 3859-3864, Dalian, Chine.

Chibelushi, C. (1997). Automatic Audio-Visual Person Recognition. Thèse de doctorat, University of Wales, Swansea.

Chiou, G., Hwang, J. (1997). Lipreading from color video, Trans. on Image Processing, vol. 6, pp. 1192-1195.

Cohen, L. (1991). On Active Contour Models and Balloons, CVGIP: Image Understanding, vol 53, pp. 211-218.

Coianiz T., Torresani L., Caprile B. (1996). 2D deformable models for visual speech analysis, D. G. Stork & M. E. Hennecke (Eds.), Speechreading by Humans and Machines: Models, Systems, and Applications pp.391-398, Springer-Verlag.

Cootes, T., Taylor, C.J., Cooper D.H., Graham J. (1992). Training Models of Shape from Sets of Examples, 3rd British Machine Vision Conference, pp. 9–18, Springer-Verlag.

Cootes, T.F., Taylor C.J., Cooper D.H. (1995). Active Shape Models - Their Training and Application, Computer Vision and Image Understanding, Vol 61, No. 1, pp. 38-59.

Cootes, T.F., Edwards G.J., Taylor C.J. (1998). Active Appearance Model, Proc. European Conference on Computer Vision, Vol 2, pp.484-498, Freiburg, Allemagne.

-
- Cootes, T.F., Walker, K.N., Taylor C.J. (2000). View-Based Active Appearance Models, Proc. 4th International Conference on Automatic Face and Gesture Recognition, pp.227–232, Grenoble, France.
- Cootes, T.F., Taylor C.J. (2001). Constrained Active Appearance Models, 8th International Conference on Computer Vision, Vol 1, pp.748–754, Vancouver, Canada.
- Cootes, T.F. (2004). Statistical Models of Appearance for Computer Vision, Technical report, free to download on <http://www.isbe.man.ac.uk/bim/refs.html>.
- Cortes, C., Vapnik, V. (1995). Support-Vector Networks, Machine Learning, Volume 20, Number 3, pp. 273-297.
- Cuavibase The CUAVE database homepage
<http://www.ece.clemson.edu/speech/cuave.htm>
- Daubias, P., Deléglise P. (2002). Statistical Lip-Appearance Models Trained Automatically Using Audio Information, EURASIP Journal on Applied Signal Processing, Vol 11, pp. 1202–1212.
- Daubias, P., Deleglise, P. (2003). The LIUM-AVS database: A Corpus to test Lip Segmentation and Speechreading systems in Natural Conditions, 8th Eur. Conf. on Speech Communication and Technology, pp. 1569-1572, Geneva, Suisse.
- Daugman, J.D, (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, pp. 1169-1179.
- De Valois, R. (1982). The orientation and direction selectivity of cells in macaque visual cortex, Vision Research, vol. 22, pp. 531-544.

Delmas, P., Coulon, P.Y. & Fristot , V. (1999). Automatic Snakes for Robust Lip Boundaries Extraction, International Conference on Acoustics, Speech and Signal Processing, vol. 6, pp. 3069-3072, Phoenix, États-Unis.

Delmas, P. (2000). Extraction des Contours de Lèvres d'un Visage Parlant par Contours Actifs, Application à la Communication Multimodale, Thèse de Doctorat, INPG (France).

Delmas, P., Eveno, N., Liévin, M. (2002). Towards Robust Lip Tracking. International Conference on Pattern Recognition, vol. 2, pp. 528-531, Québec, Canada.

Deng, G., Cahill L.W. (1995). The Logarithmic Image Processing Model and Its Applications, Proc. Of The Twenty-Seventh Asilomar Conference, Vol 2, pp. 1047-1051.

Durette, B, Héault, J. (2005). Traitement visuels biomimétiques pour la suppléance perceptive, rapport Gipsa-lab.

Eveno, N., Caplier, A., Coulon, P.Y. (2003). Jumping Snakes and Parametric Model for Lip Segmentation. International Conference on Image Processing (ICIP'03), Vol. 2, pp. 967-870, Barcelone, Espagne.

Eveno, N., Caplier, A., Coulon, P.Y. (2004). Automatic and Accurate Lip Tracking, IEEE Transactions on Circuits and Systems for Video technology, Vol. 14, N°.5, pp.706-715.

Ekman, P., Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto.

Feret, The Feret Database homepage <http://www.itl.nist.gov/iad/humanid/feret/>

Ford A., Roberts A. (1998). Color Space Conversion, Rapport technique.
<http://inforamep.net/poyton/PDFs/coloureq.pdf>.

Gacon, P., Coulon, P.Y., Bailly G. (2005). Non-Linear Active Model for Mouth Inner and Outer Contours Detection, European Signal Processing Conference (EUSIPCO'05), Antalya, Turkey.

Gacon, P. (2006). Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia. Thèse de doctorat. Institut National Polytechnique de Grenoble.

Gillespie, A. R., Kahle, A. B., Walker, R. E. (1986). Color enhancement of highly correlated images. I. Decorrelation and HIS contrast stretches, *Remote Sensing of Environment*, vol. 20, pp. 209-235.

Gong, Y., Sakauchi, M. (1995). Detection of Regions Matching Specified Chromatic Features, *Computer Vision and Image Understanding*, vol. 61, pp. 263-269.

Gordan, M., Kotropoulos, C., Pitas, I. (2001). Pseudo-automatic Lip Contour Detection Based on Edge Direction Patterns, *International Symposium on Image and Signal Processing and Analysis*, (ISPA 2001), pp. 138-143.

Granolund, G.H. (1972). Fourier Preprocessing for hand print character recognition, *IEEE Trans. Computers*, Vol C-21, pp. 195-201.

Guyader, N., Chauvin, A., Massot, C., Héraul, J., Marendaz, C. (2006). A biological model of low-level vision suitable for image analysis and cognitive visual perception, *Perception*, European Conference on Visual Perception, vol.35.

Hammal, Z., Eveno, N., Caplier, A., Coulon, P.Y. (2006). Parametric Model for Facial Features Segmentation, *Signal Processing*, vol. 86, pp. 399-413.

Hammal, Z. (2006). Segmentation des Traits du Visage, Analyse et Reconnaissance d'Expressions Faciales par le Modèle de Croyance Transférable. Thèse de doctorat. Université Joseph Fourier de Grenoble.

Hammal, Z., Couvreur, L., Caplier, A., Rombaut, M. (2007). Facial Expression Classification: An Approach based on the Fusion of Facial Deformation using the Transferable Belief Model, Int. Jour. of Approximate Reasoning, doi: 10.1016/j.ijar.2007.02.003

Harvey, L.O., Doan, V.V. (1990). Visual masking at different polar angles in the two dimensional Fourier plane, Journal of Optical Society of America A, vol. 7, pp. 116-127.

Hawken, M.J., Parker, A.J., (1987). Spatial properties of neurons in the monkey striate cortex, Proc R Soc Lond B Biol Sci., vol. 231, pp. 251-88.

Hennecke, M., Prasad, K. & Stork, D. (1994). Using deformable template to infer visual speech dynamics, Proc. 28th Annual Asilomar Conference on Signal, Systems and Computers, pp.578-582.

Hérault, J. (2001). De la rétine biologique aux circuits neuromorphiques. Traité IC2, Les Systèmes de Vision, J-M Jolion ed. Hermès.

Hsu, R.L., Abdel-Mottaleb, M., Jain, A.K. (2002). Face Detection in Color Images, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, n. 5, pp.696-706.

Hubel, D.H., Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Physiology, vol. 160, pp. 106–154.

-
- Jang, K.S., Han, S., Lee, I., Woo Y.W. (2006). Lip Localization Based on Active Shape Model and Gaussian Mixture Model, Advances in Image and Video Technology, vol. 4319, pp.1049-1058, Springer Berlin / Heidelberg.
- Jian, Y-D., Kaynak M. N., Cheok A. D., Chung K. C. (2001). Real-Time Lip Tracking for Virtual Lip Implementation in Virtual Environments and Computer Games, International Conference on Fuzzy Systems, vol. 3, pp. 1359-1362, Melbourne, Australia.
- Kanade, T., Cohn, J.F., Tian, Y. (2000). Comprehensive database for facial expression analysis, Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pp. 46-53, Grenoble, France.
- Kass, M., Witkin, A., Terzopoulos, D. (1987). Snakes: Active Contour Models, International Journal of Computer Vision, vol. 1, n. 4, pp. 321-331.
- Kaucic, R., Blake, A. (1998). Accurate, Real-Time, Unadorned Lip Tracking, International Conference on Computer Vision, pp. 370-375.
- Kolb, H., Fernandez, E., Nelson, R. (1996). Webvision : The Organization of the Retina and the Visual System. Adresse: <http://webvision.med.utah.edu/>
- Kuo, P., Hillman, P. & Hannah, J.M. (2005). Improved Lip Fitting and Tracking for Model-based Multimedia and Coding, Visual Information Engineering, pp. 251-258.
- Lallouache, T. (1991). Un Poste Visage-Parole. Acquisition et Traitement Automatique des Contours des Lèvres, Thèse de doctorat, INPG (France).

Le Goff, B., Guiard-Marigny, T., Benoît, C. (1995). Read my Lips... and my Jaws! How Intelligible are the Components of a Speaker's Face, Proc. of the European Conf. on Speech Communication and Technology, pp. 291-294, Madrid, Espagne.

Leung S.-H., Wang S.-L., Lau W.-H. (2004). Lip Image Segmentation Using Fuzzy Clustering Incorporating an Elliptic Shape Function, IEEE Transactions on Image Processing, Vol. 13, No. 1, pp. 51-62.

Li Z., Ai H. (2006). Texture-Constrained Shape Prediction for Mouth Contour Extraction and its State Estimation, Proc. ICPR06, Vol. 2, pp.88-91, Hong Kong.

Liévin, M. (2000). Analyse Entropico-Logarithmique de Séquences Vidéo Couleur, Application à la Segmentation et au Suivi de Visages Parlants, Thèse de Doctorat en Science de l'Ingénieur, INPG (France).

Liévin, M., Luthon, F. (2004). Nonlinear Color Space and Spatiotemporal MRF for Hierarchical Segmentation of Face Features in Video, IEEE Transactions on Image Processing, Vol. 13, No. 1, pp. 63-71.

Liew, A., Leung, S.H., Lau, W.H. (2000). Lip Contour Extraction using a Deformable Model, International Conference on Image Processing, Vol.2, pp. 255-258, Vancouver, Canada.

Liew, A., Leung, S.H., Lau, W.H. (2003). Segmentation of Color Lip Images by Spatial Fuzzy Clustering, IEEE Transactions on Fuzzy Systems, Vol. 11, No. 1, pp. 542-549.

Lindeberg, T. (1998). Feature detection with automatic scale selection, International Journal of Computer Vision, Vol. 30, pp. 79–116.

Lucey, S., Sridharan, S., Chandran, V. (2000). Initialised Eigenlip Estimator for Fast Lip Tracking Using Linear Regression, International Conference on Pattern Recognition, Vol.3, pp. 178-181, Barcelone, Espagne.

Luettin, J. (1997). Visual speech and speaker recognition, Thèse de doctorat, University of Sheffield, Sheffield, UK.

M2VTSbase The M2VTS database homepage <http://www.tele.ucl.ac.be/M2VTS/>

MacLeod, A., Summerfield, Q. (1987). Quantifying the Contribution of Vision to Speech Perception in Noise, British Journal of Audiology, Vol. 21, pp. 131-141.

Malciu, M. & Prêteux, F. (2000). Tracking Facial Features in Video Sequences Using a Deformable Model-based approach, Proc. SPIE Mathematical Modeling, Estimation, and Imaging, Vol. 4121, pp. 51-62.

Malague, X. (2001). Infrared and Thermal Testing, Technical Editor Xavier P.V. Malague, Editor Patrick O. Moore, American Society for Nondestructive Testing.

Mallat, S., (1999), A Wavelet Tour of Signal Processing, Academic Press, 2nd edition, ISBN: 978-0124666061.

Marcelja, S, (1980), Mathematical description of the response of simple cortical cells, Journal of Optical Society of America, Vol. 70, pp. 1297-1300.

Martinez, A.M., Benavente, R. (1998). The AR Face Database, CVC Technical Report 24.

Matthews, I., Cootes, T.F., Cox S., Harvey R., Bangham J.A. (1998). Lipreading using shape, shading and scale, Auditory-Visual Speech Processing (AVSP), pp.73-78, Australia.

Matthews, I., Baker, S. (2003). Active Appearance Models Revisited, Technical Report CMU-RITR -03-02, Carnegie Mellon University Robotics Institute.
<http://citeseer.ist.psu.edu/matthews04active.html>

McGurk, H. & McDonald, J. (1976). Hearing Lips and Seeing Voices, Nature, pp. 746-748.

Messer, K., Matas, J., Kittler, J., Jonsson, K. (1999). XM2VTSDB: The Extended M2VTS Database, In Audio- and Video-based Biometric Person Authentication, AVBPA1999.

Mirhosseini, A.R., Chen, C., Lam, K.M., Yan, H. (1997). A Hierarchical and Adaptive Deformable Model for Mouth Boundary Detection, International Conference on Image Processing, Vol. 2, pp. 756-759, Washington DC, USA.

Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K. (2002). A Coupled HMM for Audio-Visual Speech Recognition, Proc 2002 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 2, pp. 2013-2016, Orlando, USA.

Nguyen, Q. D., Milgram, M. (2008). Robust Lip Contours Localization and Tracking Using Multi Features – Statistical Shape Models. Advanced Concepts for Intelligent Vision Systems, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 5259/2008, pp. 1038-1049.

Nguyen, Q. D. (2010). Détection et suivi automatique du mouvement des lèvres. Thèse de doctorat. Université Pierre et Marie Curie.

Ojala, T., Pietikäinen, M., Mäenpää, T. (2002). Multiresolution Gray-Scale and Invariant Texture Classification with Local Binary Patterns, IEEE Transactions on Pattern analysis and Machine Intelligence, Vol. 24, No. 7, pp. 971-987.

-
- Pantic, M., Tomc, M., Rothkrantz, L.J.M. (2001). A hybrid Approach to Mouth Features Detection, Proceedings of IEEE Int'l Conf. Systems, Man and Cybernetics (SMC'01), pp. 1188-1193, Tucson, USA.
- Pardas, M., Sayrol E. (2001). Motion Estimation Based Tracking of Active Contours, Pattern Recognition Letters, Vol. 22, pp. 1447- 1456.
- Patterson, E.K., Gurbuz, S., Tufekci, Z., Gowdy J.H. (2002). Moving-Talker, Speaker-Independent Feature Study and Baseline Results Using the CUAVE Multimodal Speech Corpus, EURASIP Journal on Applied Signal Processing, Issue 11, pp.1189-1201.
- Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S. (2000). The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, pp. 1090-1104.
- Pigeon, S., Vandendorpe, L. (1997). The M2VTS multimodal face database, Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication, Eds. J. Bigun, C. Chollet, G. Borgefors, Vol. 1206, pp. 403-409.
- Poggio, T., Hulbert, A. (1998). Synthesizing a Color Algorithm from Examples, Science, Vol.239, pp.482-485.
- Potamianos, G., Neti, C., Luettin, J., Matthews, I. (2004). Audio-Visual Automatic Speech Recognition : an Overview, Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, MIT Press.
- Rabiner, L.R., Juang, B.H. (1993). Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ.

Radeva, P., Marti, E. (1995). Facial Features Segmentation by Model-Based Snakes, International Conference on Computer Vision, Massachusetts Institute of Technology, Cambridge, USA.

Rao, R., Mersereau, R. (1995). On Merging Hidden Markov Models with Deformable Templates, International Conference on Image Processing (ICIP'1995), Vol. 3, pp. 556-559, Washington DC, USA.

Rivet, B., Servière, C., Girin, L., Pham D.T., Jutten, C. (2006). Un DéTECTeur d'Activité Vocale Visuel pour Résoudre le Problème des Permutations en Séparation de Source de Parole dans un Mélange Convolutif, Journées d'Etude sur la Parole (JEP), pp. 85-88, Dinard, France.

Rosenblum, D., Saldana, M. (1996). An audiovisual test of kinematic primitives for visual speech perception, Journal of Experimental Psychology: Human Perception and Performance, Vol. 22, No. 2, pp. 318–331.

Salazar, A., Hernandez, J. E., Pietro, F. (2007). Automatic Quantitative Mouth Shape Analysis, International Conference on Computer Analysis of Images and Patterns (CAIP 2007), pp. 416-423, Vienne, Autriche.

Sadeghi, M., Kittler, J., Messer, K. (2002). Modelling and Segmentation of Lip Area in Face Images, Vision, Image and Signal Processing, Vol. 149, pp. 179-184.

Seguier, R., Cladel, N. (2003). Genetic Snakes: Application on Lipreading, International Conference on Artificial Neural Networks and Genetic Algorithms, (ICANNGA), Roanne, France.

Seo, K.H., Lee, J.J. (2003). Object tracking using adaptive color snake model, International Conference on Advanced Intelligent Mechatronics (AIM'2003), Vol. 2, pp. 1406-1410.

-
- Seyedarabi, H., Lee, W., Aghagolzadeh, A. (2006). Automatic Lip Tracking and Action Units Classification using Two-step Active Contours and Probabilistic Neural Networks, Canadian Conference on Electrical and Computer Engineering, (CCECE'2006), pp. 2021-2024, Ottawa, Canada.
- Shinchi, T., Maeda, Y., Sugahara, K., Konishi, R. (1998). Vowel recognition according to lip shapes by using neural network, The IEEE Int'l Joint Conf. on Neural Networks Proceedings and IEEE World Congress on Computational Intelligence, Vol. 3, pp. 1772-1777.
- Smirnakis, S.M., Berry, M.J., Warland, D.K., Bialek, W., Meister, M. (1997), Adaptation of Retinal Processing to Image Contrast and Spatial Scale, *Nature*, Vol. 386, pp. 69-73.
- Socolinsky, D., Selinger, A., Neuheisel, J. (2003). Face recognition with visible and thermal infrared imagery, *Comput. Vis. Image Und.*, Vol. 91, pp. 72–114.
- Stillittano, S., Caplier, A.(2008). Inner Lip Contour Segmentation by Combining Active Contours and Parametric Models, International Conference on Computer Vision Theory and Applications (VISAPP 2008), Madeira, Spain.
- Sugahara, K., Kishino, M., Konishi, R. (2000). Personal computer based real time lip reading system. International Conference on Signal Processing Proceedings, (ICSP2000), Vol. 2, pp. 1341-1346, Beijing, Chine.
- Summerfield, A.Q., MacLeod, A., McGrath, M., Brooke, M. (1989). Lips, teeth and the benefits of lipreading, *Handbook of Research on Face Processing*, A. W. Young and H. D. Ellis, Eds., pp. 223–233, Elsevier Science Publishers, Amsterdam, North Holland.

Steketee, J. (1973). Spectral Emissivity of Skin and Pericardium, *Phys. Med. Biol.*, Vol. 18, pp. 686-694.

Summerfield, A.Q. (1992). Lipreading and audio-visual speech perception, *Philosophical Transactions of the Royal Society of London, Series B*, Vol. 335, No. 1273, pp. 71–78.

Tian, Y., Kanade, T., Cohn, J. (2000). Robust Lip Tracking by Combining Shape, Color and Motion, 4th Asian Conference on Computer Vision, pp. 1040-1045, Taipei, Taiwan.

Tomasi, C., Kanade, T. (1991). Detection and Tracking of Point Features, Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, USA.

Viola, P., Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features, Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 511-518.

Vogt, M. (1996). Fast Matching of a Dynamic Lip Model to Color Video Sequences Under Regular Illumination Conditions, D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines*, Vol. 150, pp. 399–407.

Wakasugi, T., Nishiura, M., Fukui, K. (2004). Robust Lip Contour Extraction using Separability of Multi-Dimensional Distributions, Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'2004), pp. 415-420, Seoul, Korea.

Wang, S.L., Lau, W.H., Liew, A.C., Leung, S.H. (2007). Robust lip region segmentation for lip images with complex background, *Pattern Recognition*, Vol. 40, No. 12, pp. 3481-3491.

-
- Wark T., Sridharan S., Chandran V. (1998). An Approach to statistical lip modeling for speaker identification via chromatic feature extraction, Proc. 14th ICPR, Vol. 1, pp.123-125, Brisbane, Australie.
- Werda, S., Mahdi, W., Hamadou, A.B. (2007). Automatic Hybrid Approach for Lip POI Localization: Application for Lip-reading System, ICTA'07, Hammamet, Tunisia.
- Wyszecki, G., Stiles, W.S. (1982). Color Science: Concepts and Methods, Quantitative Data and Formulae, John Wiley & Sons, Inc., New York, New York, 2nd edition.
- Wojdel J.C., Rothkrantz L.J.M. (2001). Using aerial and geometric features in automatic lip-reading, Proc. 7th Eurospeech, Vol. 4, pp.2463-2466, Aalborg, Danemark.
- Wu, Z., Petar, A.Z., Katsaggelos, A.K. (2002). Lip Tracking for MPEG-4 Facial Animation, Int. Conf. on Multimodal Interfaces (ICMI'02), pp. 293-298, Pittsburgh, USA.
- Wu, Z., Aleksic, P.S. (2004). Inner lip feature extraction for MPEG-4 facial animation, International Conference on Acoustics, Speech and Signal Processing (ICASSP'2004), Vol. 3, pp. 633-636, Montreal, Canada.
- XM2VTS base The XM2VTS face database homepage,
<http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/>
- Xin, S., Ai, H. (2005). Face Alignment under Various Poses and Expressions, Affective computing and intelligent interaction, ACII 2005 First International conference, Vol. 3784, pp. 40-47, Beijing, China.
- Xu, C., Prince, J.L. (1998). Snakes, Shapes, and Gradient Vector Flow, IEEE Transactions on Image Processing, Vol.7, pp. 359-369.

Yang, J., Waibel, A. (1996). A Real-Time Face Tracker, Proc. of 3rd IEEE Workshop on Applications of Computer Vision, pp.142–147, Sarasota, USA.

Yin, L., Basu, A. (2002). Color-Based Mouth Shape Tracking for Synthesizing Realistic Facial Expressions, International Conference on Image Processing (ICIP'2002), pp. 161-164, Rochester, USA.

Yokogawa, Y., Funabiki, N., Higashino, T., Oda, M., Mori, Y. (2007). A Proposal of Improved Lip Contour Extraction Method using Deformable Template Matching and its Application to Dental Treatment, Systems and Computers in Japan, Vol. 38, pp. 80-89.

Yoshitomi, Y., Miyaura, T., Tomita, S., Kimura, S. (1997). Face identification using thermal image processing, Proc. IEEE Int. Workshop Robot Hum. Commun, pp. 374–379, Sendai, Japan.

Yuille, A., Hallinan, P., Cohen, D. (1992). Features extraction from faces using deformable templates, Int. Journal of Computer Vision, Vol. 8, No. 2, pp. 99-111.

Zahn, C.T., Roskies, R.Z. (1972). Fourier descriptors for plane close curves, IEEE Trans. Computers, Vol C-21, pp. 269-281.

Zhang, L. (1997). Estimation of the Mouth Features using Deformable Template, International Conference on Image Procesing (ICIP'1997), Vol.3, pp. 328-331, Washington DC, USA.

Zhang X., Mersereau R.M. (2000). Lip Feature Extraction Towards an Automatic Speechreading System, Proc. International Conference on Image Processing, Vol. 3, pp. 226-229, Vancouver, British Columbia, Canada.

Zhang, X., Broun, C., Mersereau, R., Clements, M. (2002). Automatic Speechreading with Applications to Human-Computer Interfaces, *Eurasip Journal on Applied Signal Processing*, Vol. 11, pp. 1228-1247.

Zhang B., Gao W., Shan S., Wang W. (2003). Constraint Shape Model Using Edge Constraint and Gabor Wavelet Based Search, *Audio-and Video-Based Biometry Person Authentication*, 4th International Conference, (AVBPA 2003), Vol. 2688, pp. 52-61, Guildford, United Kingdom.

Zhiming, W., Lianhong, C., Haizhou, A. (2002). A Dynamic Viseme Model for Personalizing a Talking Head, *International Conference on Signal Processing*, Vol. 2, pp. 1015-1018.