Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Segmentation automatique des lèvres d'un locuteur

Marc Liévin, Franck Luthon

Laboratoire des Images et des Signaux,
LIS, INPG, 46 av. Félix-Viallet, 38031 Grenoble Cedex, France
lievin@lis-viallet.inpg.fr, luthon@lis-viallet.inpg.fr
http://www- tirf.inpg.fr/PERSON/lievin/lievin.html

Résumé – Ce papier propose un algorithme non supervisé de segmentation des lèvres d'un locuteur. Une caméra fournit une séquence couleur du visage du locuteur sans éclairage ni maquillage particulier. Afin de s'affranchir des conditions d'éclairage, une transformée logarithmique définit un nouvel espace couleur (teinte, luminance). Ensuite, une approche par champ aléatoires de Markov segmente la teinte caractéristique des lèvres combinée à une détection de mouvement dans un cadre spatio-temporel. Parallèlement, la région d'intérêt principal et des paramètres géométriques sont mesurés automatiquement. Ceux-ci peuvent par la suite être intégrés dans une nouvelle étape de l'application multimédia envisagée (visiophonie, visioconférence).

Abstract – An unsupervised algorithm for speaker's lip segmentation is presented in this paper. A color video sequence of speaker's face is acquired, under natural lighting conditions and without any particular make-up. First, a logarithmic colour transform is performed from RGB to HI (hue, intensity) colour space and sequence dependant parameters are evaluated. Second, a statistical approach using Markov random field modelling segments the lip areas using red hue predominant region and motion in a spatio-temporal neighbourhood. Simultaneously, a region of interest is automatically extracted. Third, speaker's lip features are extracted to provide robust parameters for related applications (video conferencing systems).

1 Introduction

L'expression de la parole est sans nulle doute bimodale : il est par exemple reconnu que les indices visuels sont d'une aide précieuse à la reconnaissance de la parole en milieu bruité [1]. Ainsi, l'objectif premier de ce travail est de fournir des informations sur la position et le mouvement des lèvres du locuteur afin de réaliser un système automatique de reconnaissance de la parole ou de synthèse de visages parlants (Fig. 1). De nombreuses approches proposées dans ce domaine utilisent la luminance d'images statiques (*e.g.* Luettin in [6]). Les traitements classiques de segmentation et de détection dans les images à niveau de gris peuvent être alors mis en œuvre. L'utilisation de la couleur (*e.g.* Coianiz in [6]) a grandement augmenté la robustesse de la détection des lèvres mais cette approche nécessite souvent de fortes contraintes sur l'éclairage pour limiter la variabilité des paramètres de teinte. Ainsi, ceux-ci sont souvent déterminés au préalable par l'utilisateur [8].

Ici, nous présentons un algorithme de segmentation couleur des lèvres du locuteur dont les paramètres sont estimés au préalable ou séquentiellement de façon automatique. Le locuteur est actuellement équipé d'un casque muni d'une micro-caméra couleur centrée sur le bas du visage. Les images couleurs comprennent au moins la région allant des narines au menton. Le format initial des séquences est conforme à la norme PAL 768×576 sur 16 bits. L'algorithme présenté fait coopérer deux observations de teinte et de mouvement dans un cadre statistique bayésien. Les applications multimédia audiovisuelles de ce traitement sont nombreuses : visiophonie bas débit, reconnaissance automatique de la parole, agent multimédia.

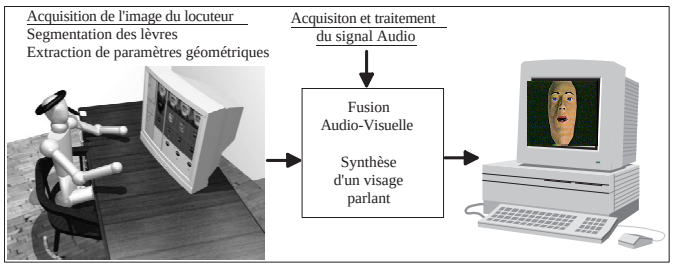

Fig. 1: Contexte de l'application : *Le Labiophone*, des paramètres pertinents sont extraits du visage du locuteur afin d'animer un clone de synthèse.

2 Approche Couleur

La première étape de la transformée couleur des images de la caméra convertit le format vidéo Y/C en RVB, espace couleur fréquemment employé en traitement d'image (Eq. 1).

$$\begin{bmatrix} R \\ V \\ B \end{bmatrix} = \begin{bmatrix} 1 & 1.402 & 0 \\ 1 & -0.714 & -0.334 \\ 1 & 0 & 1.772 \end{bmatrix} * \begin{bmatrix} Y \\ Cr - 128 \\ Cb - 128 \end{bmatrix} \quad (1)$$

L'étude de la distribution des vecteurs couleurs de l'image d'un locuteur sous éclairage naturel fait apparaître dans le cas d'une caméra mono-CCD des corrélations et des orientations privilégiées (Fig. 2).

Ainsi, les lèvres ont une teinte caractéristique (plutôt rouge) qui justifie l'approche par teinte [7]. La conversion

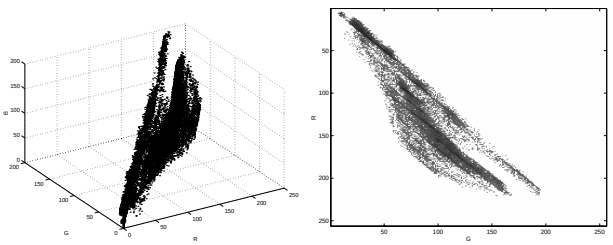

FIG. 2: *A gauche :* Histogramme RVB d'un visage sans éclairage d'appoint; *A droite :* Histogramme chromatique RV de ce même visage.

la plus courante est celle vers l'espace HIS (teinte, luminance, saturation), qui tente de différencier l'attribut de la couleur (teinte) de la norme de la radiation (intensité). Cette approche, même par transformée linéaire, fait intervenir des rapports de différences dans son expression qui rendent les calculs très sensibles au bruit et donc peu robustes aux conditions de prise de vue. A l'inverse, les modèles d'éclairement les plus simples montrent qu'une approche logarithmique permet de s'affranchir dans les calculs d'intensité des coefficients de réflexion des sources lumineuses et par conséquent des variations d'éclairement dues à l'image observée. Ainsi, notre algorithme combine les couleurs RVB de façon logarithmique afin d'obtenir une évaluation de la teinte indépendante des conditions d'éclairement. Pour cela, l'évaluation de cette teinte s'effectue dans le cadre du modèle LIP (Logarithmic Image Processing) développé par l'équipe de Jourlin [3]. Le développement au premier ordre de la différence logarithmique s'exprime alors comme le rapport des composantes vertes et rouges (Eq. 2). On obtient ainsi les deux composantes du nouvel espace couleur HI (Fig. 3).

$$H = 256 \times \frac{V}{R} \qquad \text{et} \qquad I = \frac{R+V+B}{3} \tag{2}$$

FIG. 3: *En haut*: 5 images consécutives de luminance I; *En bas* : les images de teinte H correspondantes.

3 Segmentation des lèvres

3.1 Teinte et mouvement

Deux observations o sur l'espace couleur HI ont été retenues comme les plus pertinentes : une fonction de présence de la teinte des lèvres h (hue) et une information de mouvement fd (frame difference) (Fig. 4).

La première observation correspond à un filtrage de la teinte (parabolique ou gaussien) autour d'une valeur particulière (Eq. 3); la seconde est un gradient temporel simplifié sur l'intensité (Eq. 4).

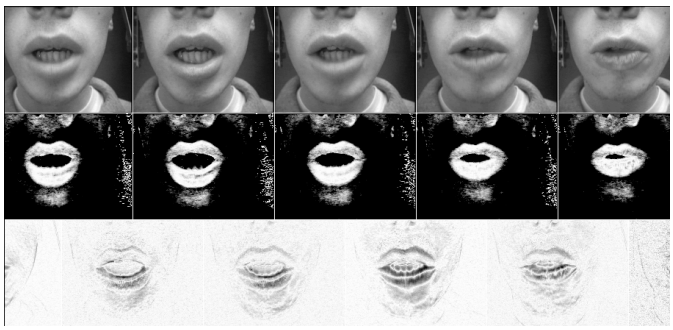

FIG. 4: *De haut en bas :* Luminance de la séquence initiale; Observation sur la teinte des lèvres (en blanc); Observation de mouvement sur la luminance (en noir).

$$h(s) = \left[256 - \left(\frac{H(s) - H_m}{\Delta_H}\right)^2\right] \times 1_{\frac{|H(s)-H_m|}{\Delta_H} \leq 16} \tag{3}$$

$$fd(s) = |I_t(s) - I_{t-1}(s)| \tag{4}$$

Les valeurs de seuils et les paramètres de ces observations sont estimés automatiquement au cours du traitement : la dispersion de la teinte des lèvres et sa valeur moyenne sont estimées par raffinement d'histogramme; le seuil de bruit de la caméra visible sur le gradient temporel est estimé par un calcul d'entropie.

3.2 Etiquettes et régularisation

Les observations sont seuillées, produisant des champs initiaux comportant les informations de teinte et de mouvement. Ces champs initiaux sont bruités et incomplets. Afin de les préciser, on met en œuvre une régularisation statistique s'appuyant sur une modélisation par champ de Markov. Dans cette perspective, la segmentation de la région des lèvres nécessite un étiquetage de chaque site pixel. L'ensemble des configurations binaires, combinant les informations élémentaires de teinte et de mouvement, est donné par le tableau 1.

TAB. 1: Tableau des 4 étiquettes codant les 2 informations élémentaires de teinte et de mouvement

Teinte	0	0	1	1
Mouvement	0	1	0	1
Etiquettes	b_0	b_1	a_0	a_1

Considérant l'équivalence entre champ de Markov et distributions de Gibbs, le critère du maximum a posteriori appliqué aux champ de Markov permet de maximiser la probabilité d'attribution du champ d'étiquettes E relativement aux observations O en minimisant une énergie globale W sur un voisinage η du site pixel considéré s (Fig. 5) [2].

L'expression de W fait intervenir deux énergies : la première U_o exprime l'attache aux données et tend à minimiser le bruit sur les observations, la seconde U_m représente l'énergie associée au modèle a priori et contraint la configuration spatio-temporelle de l'étiquetage au cours de la relaxation.

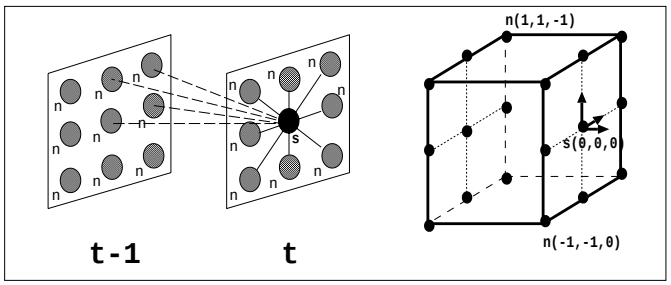

FIG. 5: *A gauche*: Voisinage spatio-temporel η considéré avec s pour site-pixel (en noir) et n un voisin (en gris); *A droite*: le cube métrique élémentaire C_{xyt} correspondant.

$$W(S) = \sum_{o\in\{fd,h\}} U_o(S) + \lambda.U_m(S) \tag{5}$$

L'énergie U_o résulte de l'adéquation aux observations définie par $o(s) = \psi_o(e_s) + n_o$, n_o bruit gaussien centré de variance σ_o^2 (Eq. 6). La fonction d'adéquation ψ_o est calculée de façon à respecter le critère du bruit additif centré.

$$U_o(S) \quad = \quad \sum_{s\in S}\left[\frac{[o_s - \psi_o(l_s)]^2}{2\sigma_o^2}\right] \tag{6}$$

L'énergie a priori U_m définit les contraintes d'homogénéité spatiales et temporelles du modèle en intégrant des fonctions potentielles locales $V(e_s, e_n)$ (Eq. 7).

$$U_m(S) \quad = \quad \sum_{s\in S}\Big[\sum_{n\in\eta(s)} V(e_n, e_s)\Big] \tag{7}$$

Par analogie avec les forces électrostatiques décrites en physique, la fonction d'interaction entre sites V est définie comme inverse de la distance euclidienne entre deux voisins dans le cube métrique C_{xyt} (Eq. 8). Les facteurs d'échelle β_x, β_y et β_t favorisent l'homogénéité spatiale et temporelle en l'absence de mouvement. On définit par ailleurs $\beta_x = 2.\beta_y = \beta_s$ afin de favoriser les horizontales en raison de la forme étirée des lèvres et de ses mouvements principalement verticaux (Eq. 9).

$$V \quad = \quad \frac{1}{\sqrt{\left(\frac{\delta_x}{\beta_x}\right)^2 + \left(\frac{\delta_y}{\beta_y}\right)^2 + \left(\frac{\delta_t}{\beta_t}\right)^2}} \tag{8}$$

$$= \quad \frac{\beta_s\beta_t}{\sqrt{\beta_t^2\left(\delta_x^2 + 4\delta_y^2\right) + \beta_s^2\delta_t^2}} \tag{9}$$

3.3 Résultats de segmentation

Le minimum d'énergie pour chaque site pixel est obtenu par un algorithme de relaxation déterministe itératif (ICM: Iterated Conditional Modes), chaque étiquette possible est envisagée pour chaque site. Le champ d'étiquettes initial E_t^0 est celui obtenu par seuillage estimé sur les observations. Après quelques itérations (en moyenne de 5 à 10), on obtient des masques de teinte et de mouvement homogènes.

Par ailleurs, l'algorithme a la possibilité d'extraire dynamiquement la région d'intérêt et de n'effectuer la relaxation que sur cette région. Une meilleure localisation de la région de traitement procure une plus grande précision sur les calculs d'adéquation et réduit le temps de calcul d'un facteur non négligeable.

On obtient au final des masques robustes aux conditions d'éclairage et indépendants du locuteur (Fig. 6).

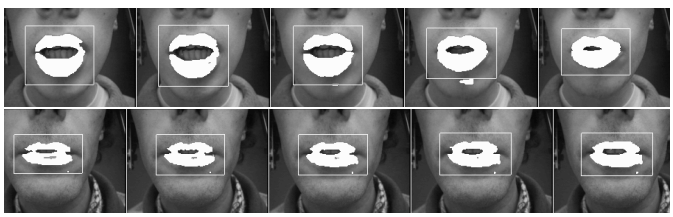

FIG. 6: Résultats de la segmentation sur deux locuteurs sans éclairage particulier, le masque des lèvres et la région d'intérêt sont superposés à la luminance.

4 Exploitation des résultats

4.1 Paramètres labiaux

Dans la perspective d'une coopération audiovisuelle dans le suivi des lèvres, les paramètres labiaux les plus pertinents sont ceux référant au conduit vocal. L'étirement et l'écartement interne sont donc deux mesures importantes du suivi des lèvres. Le réglage manuel des paramètres de segmentation procurait une grande précision sur les masques finaux [5]. La figure 7 montre la précision obtenue avec notre algorithme dans le cas automatique et non supervisé de la mesure de l'étirement interne B. On remarque que l'erreur de mesure avec la vérité terrain n'excède pas un pixel.

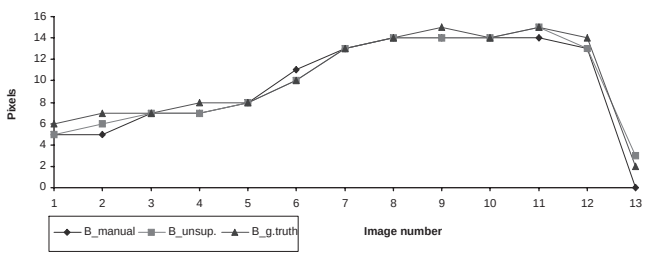

FIG. 7: Exemple de mesure de l'étirement interne des lèvres suivant trois méthodes sur une séquence d'images: B_{manual} mesure avec réglage manuel des paramètres de segmentation; $B_{unsup.}$ mesure automatique avec estimation des paramètres de segmentation; $B_{g.truth}$ mesure manuelle vérité terrain

4.2 Paramètres géométriques et sémantiques

Si l'application envisagée est la synthèse d'un visage parlant, l'interpolation B-spline nécessite de nombreux points de contrôle. La figure 8 montre un exemple d'ex-

traction des points de contour du masque des lèvres à partir desquels un modèle global peut être appliqué (descripteurs de Fourier, contours actifs). La figure 9 montre le résultat par simple chaînage de ces points.

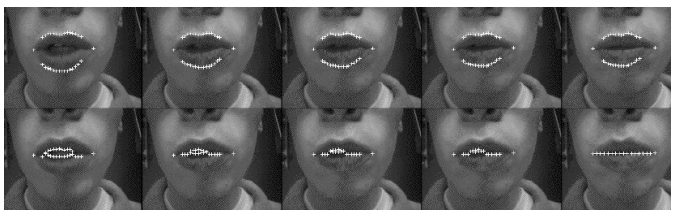

FIG. 8: Exemple d'extraction de contours internes et externes du masque des lèvres.

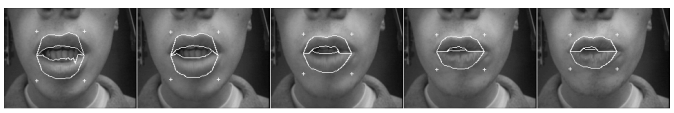

FIG. 9: Exemple de chaînage des points de contours.

Des informations sémantiques (e.g. ouverture/fermeture) sont également évaluées sur le masque final segmenté. De telles informations donnent une indication non négligeable sur la position de la bouche et par conséquent fournissent une aide à son suivi et sa représentation.

5 Conclusion et perspectives

Les résultats actuels de notre algorithme mettent en évidence des aspects essentiels pour le traitement des lèvres et plus généralement des visages parlants. Tout d'abord, la couleur permet de segmenter efficacement et avec précision, à condition d'appliquer un changement d'espace bien approprié aux conditions de prise de vue. Ensuite, l'apport de l'information temporelle permet de préciser la segmentation pour des conditions difficiles. De même, l'évaluation dynamique de la zone d'intérêt réduit le temps de calcul et augmente la qualité des masques obtenus. Dans ce contexte, la régularisation par champ de Markov propose un cadre théorique statistique permettant d'intégrer ces notions dans un voisinage spatio-temporel.

Les développements actuels s'orientent vers une approche par caméra portable de bureau et une coopération régions-contours afin de préciser les formes caractéristiques des lèvres telles que les commissures ou l'arc de Cupidon. Dans le cadre du projet Labiophone (Fédération Elesa n°8 (CNRS/INPG)), notre algorithme se situe au premier niveau de la chaîne de traitement et fournit la localisation et les points caractéristiques des lèvres. Cette mesure est utilisée pour l'évaluation des paramètres géométriques des lèvres par optimisation d'un algorithme de contours actifs (Fig. 10) [4] et devient une étape indispensable à la suite du traitement. Le temps de calcul de cette première étape est de l'ordre de quelques secondes par image 256×256 pixels sur station de travail 150 MHz. La mise en œuvre sur processeur vidéo est par conséquent envisagée.

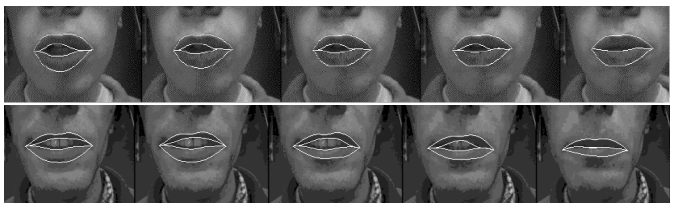

FIG. 10: Résultats du suivi du contour des lèvres par contours actifs.

En conclusion, la qualité et la robustesse aux conditions de prises de vues des premiers résultats intégrant les contours actifs assure pour la suite des mesures géométriques fiables autorisant des applications du type visio-conférence ou visiophone haut de gamme.

Remerciements

Les auteurs remercient P. Delmas et P.Y. Coulon pour leur collaboration dans le développement de ce travail (résultats de suivi par contours actifs Fig. 10).

Références

[1] C. Benoît, M.T. Lallouache, and T. Mohamadi. A set of French visemes for visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, pages 485–504. Elseviers Science Publishers, 1992.

[2] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Int.*, 6(6):721–741, November 1984.

[3] M. Jourlin and J-C. Pinoli. Image dynamic range enhancement and stabilization in the context of the logarithmic image processing model. *Signal Processing*, 41(2):225–237, January 1995.

[4] M. Liévin and P. Delmas et al. Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme. In *Proc. of IEEE Int. Conf. on Multimedia, Computing and Systems*, Florence, Italie, June 1999.

[5] M. Liévin and F. Luthon. Lip features automatic extraction. In *Proc. of IEEE Int. Conf. on Image Processing*, volume 3, pages 168–172, Chicago, Illinois, October 1998.

[6] D. Stork and M. Hennecke. *Speechreading by Humans and Machines*, volume 150. Springer-Verlag, Berlin, 1996.

[7] M. Vogt. Interpreted multi-state lip models for audio-video speech recognition. In *Proceedings of the Audio-Visual Speech Processing, Cognitive and Computational Approaches Workshop*, Rhodes (Greece), September 1997.

[8] T. Wark and S. Sridharan. A synthetic approach to automatic lip feature extraction for speaker identification. In *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pages 3693–3696, Seattle, Washington, USA, May 1997.