VERS L'EXTRACTION AUTOMATIQUE DES LÈVRES D'UN VISAGE PARLANT

Nicolas EVENO [1], Patrice DELMAS [2], Pierre-Yves COULON [3]

[1,3] Laboratoire Des Images et Signaux
LIS, INPG, 46 av. Félix Viallet, 38031 Grenoble Cedex, France

[2] Centre for Image Technology and Robotics
Tamaki Campus, The University of Auckland, Private Bag 92019, Auckland, New Zealand

eveno,coulon@lis.inpg.fr
patrice@cs.auckland.ac.nz

Résumé - Cet article présente un algorithme visant à extraire de façon automatique les contours labiaux d'un locuteur dans une séquence vidéo sans contrainte d'éclairage ou de maquillage. Il fait partie d'un projet dont le but est la création d'un outil de communication audiovisuelle bas débit (projet RNRT/TEMPOVALSE). Cet algorithme utilise les informations spatiales (approche région/contour) et temporelles (fonction de similarité) des plans Luminance et Teinte pour la convergence rapide et robuste de contours actifs vers les contours des lèvres. L'association de l'algorithme de suivi de Kanade-Lucas et de notre méthode d'extraction de points caractéristiques de la bouche assure un positionnement automatique, rapide et robuste des snakes. Cela permet d'augmenter leur fiabilité tout en approchant une cadence de traitement temps réel.

Abstract *- An algorithm for speaker's lip automatic extraction in video sequences is presented here. This work is a part of the TEMPOVALSE project, an advanced audio-visual communication tool which aims at providing a very low bit rate coding. Our method uses spatial (region and contour) and temporal (similarity fonction) informations from Luminance and Hue components. This allows fast and accurate convergence of active contours towards lips boundaries. The use of Kanade-Lucas tracking algorithm with our point extraction method ensures an automatic, fast and robust initialisation of snakes. The significant robustness boost and related computationnal cost enhancement allows us to approach real time processing.*

1. Introduction

Il a été prouvé que les informations visuelles propres à la locution améliorent significativement la reconnaissance de la parole en environnement bruité [6]. L'apparition de la langue ainsi que l'ouverture et la position des lèvres sont des indices que le cerveau utilise efficacement pour améliorer l'intelligibilité du discours. L'objectif du travail présenté ici est l'extraction automatique de ces indices sur des visages non maquillés et en lumière naturelle. A terme, le système doit permettre de créer un avatar de synthèse du locuteur pour des applications multimedia du type visiophonie ou interface homme-machine (voir figure 1). De plus, la cadence de traitement doit être rapide (environ 15 images par seconde) afin d'assurer une intelligibilité de la parole suffisante.

Nous proposons ici un algorithme d'extraction des contours des lèvres robuste vis-a-vis du changement de locuteur et des conditions d'éclairage. Le locuteur est équipé d'un casque muni d'une micro-caméra couleur centrée sur le bas du visage. Les images comprennent au moins la région allant des narines au menton. L'algorithme présenté intègre une phase de pré-segmentation permettant de localiser précisément les points caractéristiques de la bouche [4] (voir figure 2). La phase de détection et de suivi des contours labiaux est assurée par les contours actifs [5] couplés à l'algorithme de suivi de points de Kanade-Lucas.

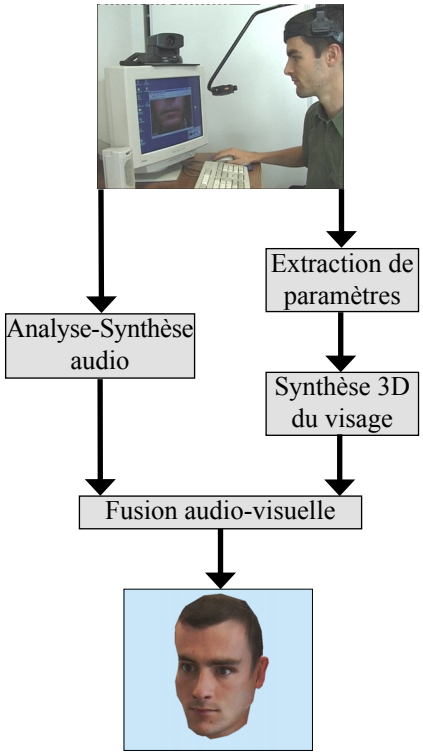

Figure 1: Contexte de l'application : Le Labiophone (projet Tempovalse); des paramètres pertinents sont extraits du visage du locuteur an d'animer un clone de synthèse.

2. Extraction des points caractéristiques des lèvres

Dans le cas particulier de l'extraction des contours labiaux, la localisation des points caractéristiques des lèvres (commissures et extrema verticaux de la bouche) est une étape essentielle pour une détermination de contours précis [7].

La détermination des commissures de la bouche se fait par une méthode proposée par P. Delmas [4]. On peut observer que les zones les plus sombres de l'image sont localisées au niveau des commissures de la bouche ainsi qu'à l'intérieur de celle-ci, qu'elle soit ouverte ou fermée. Il paraît donc intéressant de déterminer pour chaque colonne de l'image la position du minimum de luminance (voir figure 2-a). Afin de tenir compte de l'aspect symétrique et du centrage horizontal de la bouche, on introduit une fonction de pondération favorisant les minima proches du centre de l'image plutot que ceux situés sur les bords, à priori en dehors de la bouche (voir figure 2-c) :

$$\varsigma_j = e^{-4\left(1-2\frac{j}{N_{col}}\right)^2} \qquad (1)$$

où N_{col} est le nombre de colonnes de l'image. On construit alors un vecteur d'accumulation noté V_{roi} , somme des projections pondérées des minima précédemment détectés (voir figure 2-b). Considérons que, pour une colonne j, je minimum de luminance se situe à la ligne i. La composante i du vecteur V_{roi} est alors incrémentée d'une valeur ζ_j correspondant à la pondération de la colonne j. La composante la plus forte de V_{roi} donne alors la position de la bouche selon l'axe vertical. Pour trouver les commissures, on effectue ensuite un chaînage des minima de luminance (voir figure 2-d).

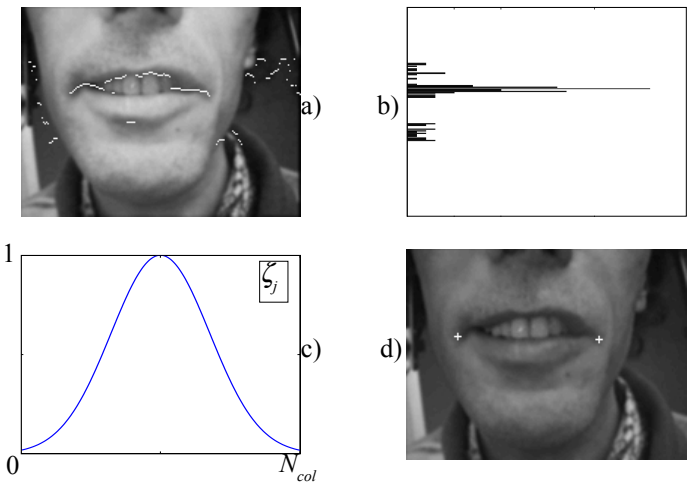

FIG. 2 : (a) : minima de luminance par colonnes ; (b) : vecteur V_{roi} ; (c) fonction de pondération ζ_j ; (d) : commissures extraites.

La détermination des 4 extrema verticaux de la bouche est ensuite effectuée par une approche région/contour appliquée à un modèle constitué de paraboles et de quartiques (figure 3-a). Les commissures du modèle sont mises en coïncidence avec celles qui ont été estimées par chaînage, puis il est déformé verticalement. Il s'agit de minimiser une fonction qui intègre les informations de gradient sur son contour et les informations de teinte sur sa zone intérieure. On obtient ainsi la position optimale des 4 extrema verticaux de la bouche [1] (figure 3-b).

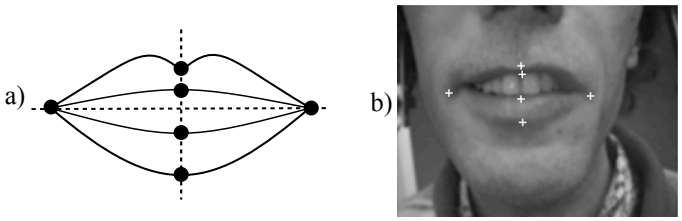

FIG. 3 : (a) : Modèle de bouche ; (b) : Points caractéristiques des lèvres extraits.

3. Extraction et suivi des contours labiaux

3.1 Suivi des points caractéristiques

La méthode décrite en section 2 permet d'extraire les points caractéristiques de la bouche dans l'image initiale. Leur position dans les images suivantes est obtenue par l'algorithme de Kanade-Lucas [9]. Dans ce cas, seuls les voisinages des points caractéristiques sont traités, ce qui apporte un gain de temps significatif par rapport à la méthode d'extraction directe (section 2).

On suppose que le voisinage du point suivi dans l'image I se retrouve dans l'image suivante J par une translation :

$$J(x,y) = I(x-\alpha, y-\beta) + n(x,y) \qquad (2)$$

où $(\alpha,\beta)^T$, qui sera désormais noté $\boldsymbol{d}$, est le vecteur déplacement et $n(x,y)$ est le bruit. $I(x,y)$ et $J(x,y)$ sont des scalaires qui peuvent être, par exemple, la valeur de la luminance au point de coordonnées (x,y). Pour notre application, il semble que le plan rouge donne de meilleurs résultats. La figure 4 illustre l'égalité (2) dans le cas d'un signal mono dimensionnel.

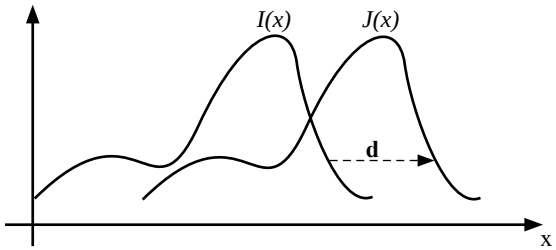

FIG. 4 : Un voisinage dans l'image I se retrouve dans l'image suivante J par une translation de vecteur $\boldsymbol{d}$.

Le vecteur $\boldsymbol{d}$ est alors choisi de manière à minimiser l'erreur ε, calculée sur la fenêtre de voisinage $\mathcal{W}$ de la manière suivante :

$$\varepsilon = \iint_W [I(\mathbf{x}-\mathbf{d}) - J(\mathbf{x})]^2 \, \omega(\mathbf{x}) \, d\mathbf{x} \qquad (3)$$

où $\omega(\boldsymbol{x})$ est une fonction de pondération et $\boldsymbol{x} = (x,y)^T$.

En général, $\omega(\mathbf{x})$ est constante et vaut 1. Mais elle peut également prendre une forme gaussienne si on veut donner plus d'importance au centre de la fenêtre.

La résolution de l'équation (3) (détaillée dans [11]) conduit à :

$$G\,\mathbf{d} = \mathbf{e} \tag{4}$$

où :

$$\begin{cases} G = \iint_W \mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x})\omega(\mathbf{x})\,d\mathbf{x} \\ \mathbf{e} = \iint_W (I(\mathbf{x}) - J(\mathbf{x}))\mathbf{g}(\mathbf{x})\omega(\mathbf{x})\,d\mathbf{x} \\ \mathbf{g}^T = \left(\dfrac{\partial I(\mathbf{x})}{\partial x} \quad \dfrac{\partial I(\mathbf{x})}{\partial y} \right) \end{cases}$$

L'équation (4) permet de déterminer le vecteur déplacement $\boldsymbol{d}$ et donc d'estimer la position du point suivi dans l'image J.

Durant le suivi, on suppose que le voisinage des points caractéristiques subit une simple translation. En toute rigueur, il peut également subir une déformation. Des modèles affines incluant une matrice de déformation ont également été développés [10]. Mais, dans le cas d'images consécutives d'une séquence vidéo la déformation de la fenêtre W est faible. Dès lors, l'utilisation du modèle affine ralentit considérablement les calculs sans apporter de gain significatif à la précision de l'estimation. De plus, de manière à minimiser les déformation subies par les voisinages des points caractéristiques, les fenêtres d'analyse ne sont pas toujours centrées sur les points à suivre (figure 5). Elles sont décalées de manière à ne contenir que des zones peu déformables. Ainsi, les voisinages des points intérieurs et des commissures n'empiètent pas sur l'intérieur de la bouche car, lors de l'ouverture ou de la fermeture, celui-ci peut évoluer rapidement : les dents peuvent apparaître subitement et les lèvres peuvent se superposer. Dès lors, les calculs de similitude d'une image à l'autre sur cette zone ne peuvent aboutir.

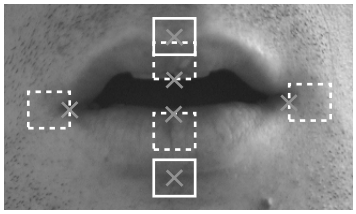

FIG. 5 : voisinages utilisés pour effectuer le suivi des points caractéristiques (repérés par des croix). Les voisinages des commissures et des points intérieurs (en pointillés) sont décalés de manière à ne pas empiéter sur l'intérieur de la bouche.

La figure 6 présente les résultats obtenus sur une séquence vidéo en utilisant des fenêtres d'analyse carrées de taille 21×21. On peut constater que l'algorithme de Kanade-Lucas permet la plupart du temps de suivre correctement les points caractéristiques (figure 6-a). Cependant, si les mouvements des lèvres sont rapides, le suivi des points intérieurs peut s'avérer difficile (voir figure 6-b). Le calcul de l'erreur ε pour chaque point permet de déterminer l'exactitude de l'estimation. Si l'erreur est trop importante, le point est repositionné en utilisant la technique décrite en section 2. Cette nouvelle initialisation est beaucoup plus rapide que la première car elle ne concerne en général qu'un ou deux points. Le modèle est donc ajusté en ne déformant qu'une ou deux paraboles.

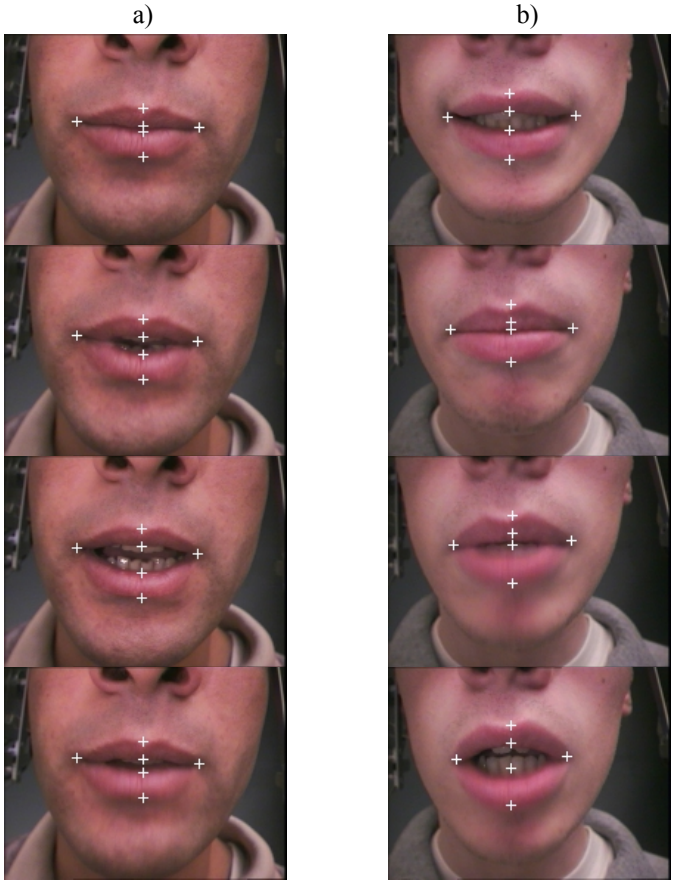

FIG. 6 : résultats obtenus sur des séquences vidéo en utilisant des fenêtres d'analyse carrées de taille 21×21. (a) : tous les points sont bien suivis. (b) : les points intérieurs sont mal suivi car le mouvement des lèvres est rapide.

3.2 Détermination des contours labiaux

Classiquement, les méthodes de détection de contours combinent une détection globale des points contours et un processus de chaînage (souvent basé sur des informations locales de l'image). Pour leur capacité à intégrer ces deux étapes en une seule, nous avons choisi les contours actifs. Les contours actifs ou «snakes» [5] recherchent des contours chaînés évoluant, à partir d'une forme initiale prédéfinie, sous l'effet d'une méthode d'optimisation (de type «descente de gradient») en utilisant les données images (généralement une carte des niveaux de gris ou du gradient) propre aux formes recherchées.

La forme utilisée pour l'initialisation des «snakes» est le modèle présenté à la figure 3-a. Ses paramètres sont fixés à chaque image par la position des points caractéristiques. Pour l'image initiale, l'extraction se fait par la méthode décrite en section 2. Pour les images suivantes, la position des points est obtenue par l'algorithme de suivi de Kanade-Lucas [9]. Si la bouche est ouverte, les contours intérieur et extérieur sont tous

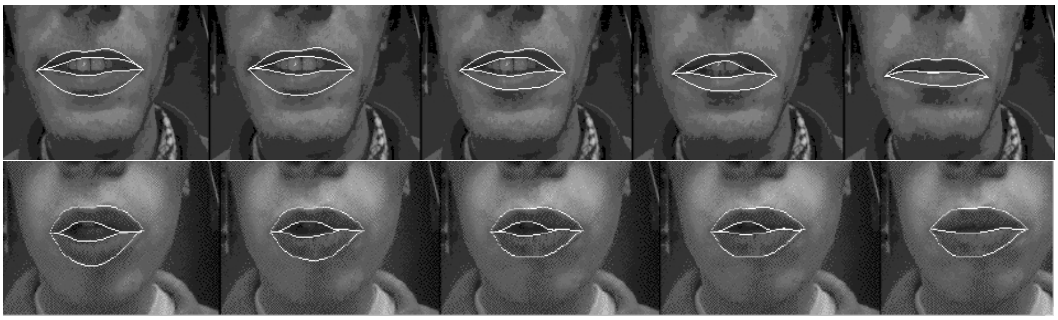

FIG 7 : Détection et suivi des contours labiaux sur deux séquences d'images.

les deux déterminés par les contours actifs. Si elle est ouverte, le contour intérieur se réduit à une ligne et est déterminé par chaînage des minima de luminance entre les commissures.

En ne traitant que les voisinages d'un nombre réduit de points pour l'initialisation, cet algorithme permet un gain de temps significatif par rapport à la méthode décrite précédemment (section 2) tout en conservant une bonne précision pour l'estimation des points considérés. Les contours actifs sont dès lors beaucoup moins sensibles aux réglages de leurs paramètres et convergent rapidement. La convergence est assurée par l'utilisation d'un critère de distance quadratique et/ou du calcul d'un critère de confiance basé sur les valeurs du gradient dans une bande étroite le long du contour actif. Une nouvelle méthode d'inversion littérale de la matrice de rigidité des contours actifs [4] ainsi que l'utilisation d'instructions MMX et SIMD [3] permettent d'approcher la cadence temps réel.

On obtient au final des contours labiaux précis, robustes aux conditions d'éclairage et indépendants du locuteur (voir figure 7).

4. Perspectives

Les développements actuels s'orientent vers une approche par web-cam et une coopération régions/contours afin de préciser les formes caractéristiques des lèvres telles que les commissures ou l'arc de cupidon et d'obtenir des informations sémantiques précises (bouche en ouverture, en fermeture) [8]. S'affranchir d'une image trop cadrée est un des axes de recherche actuelle notamment par l'utilisation de formes elliptiques et d'approche Teinte/Luminance [1].

5. Références

[1] S. Bircheld. "Elliptical head tracking using intensity gradients and color histograms". *IEEE Conf. Computer Vision and Pattern Recognition, CVPR,* Santa Barbara, California*,* 1998.

[2] T. Chen and R.R. Rao, "Audio-visual integration in multimodal communication". *Special Issue on Multimedia Signal Processing, IEEE Proceedings*, pp. 837-852, 1998.

[3] H. Cruchon. Projet ingenieur : "Implantation et optimisation d'un algorithme de contours actifs". *Technical report,* Laboratoire des Images et des Signaux de Grenoble, Grenoble, 2000.

[4] P. Delmas. "Extraction des contours des lèvres d'un visage parlant par contours actifs. Application à la communication multimodale". Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 2000.

[5] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contours models". *International Journal of Computer Vision*, pages 321-331, 1988.

[6] B. LeGoff, T. Guiard-Marigny, M. Cohen, and C. Benoit. "Real-time analysis-synthesis and intelligibility of talking faces". XXeme Journees d'Etude sur la Parole, 1994.

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

[8] M. Lievin. "Analyse entropico-logarithmique de séquences vidéo couleur. Application à la segmentation et au suivi de visages parlants". Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 2000.

[9] B. D. Lucas and T. Kanade. "An iterative image registration technique with an application to stereo vision". *Proc. on International Joint Conference on Artificial Intelligence, IJCAI*, Vancouver, 1981.

[10] J. Shi and C. Tomasi. "Good features to track". *IEEE Conference Computer Vision and Pattern Recognition, CVPR*, Seattle, 1994.

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