Extraction temps-réel de contours labiaux par segmentation vidéo robuste en vue d'animation 3D

Brice BEAUMESNIL, Franck LUTHON, Marc CHAUMONT

Laboratoire LIUPPA, IUT Informatique,
Château-Neuf, Place Paul Bert, 64100 Bayonne, France
beaumesn@iutbayonne.univ-pau.fr, Franck.Luthon@univ-pau.fr,
Marc.Chaumont@iutbayonne.univ-pau.fr

Résumé – Nous présentons une chaîne complète analyse/synthèse temps réel permettant l'animation labiale d'un clone de synthèse à partir d'une simple caméra non calibrée dans un environnment non contraint (typ. caméra motorisée ou webcam avec des conditions d'éclairage de bureau). La technique utilisée est basée sur une segmentation labiale à partir d'une teinte chair extraite d'un espace couleur non-linéaire robuste aux changements de luminosité. Les contours internes et externes sont ensuite extraits puis interprétés afin d'animer en temps-réel et de façon réaliste la bouche d'un clone de synthèse.

Abstract – We present a complete real-time analysis/synthesis approach allowing lip animation of a clone with a single camera in unconstrained environment (typ. motorized camera or webcam). This approach is based on a labial segmentation of a skin hue computed from a non-linear color space that is robust to lighting variations. Internal and external contours are extracted and interpreted to make a real time realistic animation of a mouth's clone.

1 Introduction

Plus de 80% de l'information visuelle lors d'une conversation entre deux personnes provient des mouvements de la bouche. Ces indices visuels sont essentiels à une meilleure compréhension de la parole [1]. Ainsi dans le cadre d'applications nécessitant l'animation d'un clone, l'effort porte sur la reconstruction réaliste des mouvements des lèvres.

L'objectif de ce papier est de faire le lien entre l'analyse de visage parlant et l'animation de clone de synthèse, afin d'animer en temps réel celui-ci. Nous nous intéressons à l'analyse labiale du locuteur pour l'animation de la bouche d'un clone uniquement à partir de la vidéo sans utiliser l'information sonore. L'important n'est pas d'avoir des résultats d'analyse très précis mais plutôt d'avoir un rendu de synthèse acceptable (dans notre cas une animation réaliste). Ainsi, pour l'application visée, l'effort en terme de complexité opératoire n'est pas à porter sur un algorithme d'analyse particulier ni sur un modèle de synthèse mais sur la chaîne de traitement globale temps-réel entre la vidéo (en entrée de chaîne) et le clone (en sortie de chaîne).

Les algorithmes développés actuellement par la communauté reposent sur des méthodes de haut-niveau, utilisant des bases d'apprentissage de grande taille permettant la convergence vers une situation connue ou estimée [2]. La chaîne de traitement que nous proposons ici se situe en marge de ces méthodes : nous cherchons à utiliser des algorithmes de bas-niveau très rapides et sans connaissance du visage, capables de s'adapter à n'importe quelle forme de bouche.

Nous pouvons diviser notre algorithme en trois étapes :

– La première étape consiste à effectuer une segmentation couleur de bas-niveau sur un espace peu sensible aux variations de luminosité et révélant plus particulièrement les zones de teinte chair.
– La deuxième étape utilise une approche basée sur les contours actifs pour délimiter les contours internes et externes des lèvres.
– Enfin la dernière étape utilise l'information des contours des lèvres pour animer le clone.

2 Description de la chaîne de traitement

Notre étude se situe en aval du développement d'un outil de détection et de suivi de visage permettant un cadrage dynamique optimal de celui-ci par asservissement de caméra. Nous plaçons donc, dans ce papier, directement le locuteur en face de la caméra de sorte que son visage occupe la majorité de l'image (l'image est ainsi considérée comme le cadre de recherche). Par conséquent, le locuteur devra rester toujours à peu près à la même distance de la caméra lors d'une acquisition, afin que les proportions de bouche restent les-mêmes (étant donné que le cadrage est statique).

Enfin nous demandons seulement au locuteur d'avoir une position neutre (c'est-à-dire bouche fermée) sur la première image afin d'initialiser les proportions de la bouche (hauteur et largeur) pour l'animation du clone (cela permet de prendre en compte le recul du locuteur face à la caméra, qui peut être différent selon les acquisitions).

Les conditions d'éclairage ne sont pas stabilisées afin d'être dans les conditions les plus réalistes possibles (ajout ou suppression de sources lumineuses, éclairage non uniforme, ...).

Notre algorithme de segmentation a été réalisé sur la base

de deux travaux : une approche de classification markovienne proposée par Liévin [3], et une approche contours actifs présentée par Delmas [4]. Notre démarche a consisté à adapter ces méthodes pour atteindre trois objectifs :

- fonctionner dans un environnement non contraint (conditions d'éclairage et mouvements du visage non stabilisés)
- être exécutable en temps-réel
- coupler analyse et synthèse pour effectuer une animation labiale réaliste d'un clone de locuteur.

2.1 Extraction de la zone labiale

Afin de segmenter le visage pour détecter la zone labiale, nous utilisons l'espace couleur LUX [3]. L'apport principal de cet espace couleur, non linéaire par rapport à l'espace couleur RGB, est d'amplifier le contraste tout en s'affranchissant le plus possible des variations d'éclairage. De plus, l'information principale d'un visage humain est représentée, dans cet espace, par la composante U (teinte rouge) dans le cas particulier où $R > L$ (L représentant la luminance). Ce qui nous permet de définir un modèle discriminant de visage grâce à la teinte que l'on nomme H (Eq.2) modèle sur lequel nous nous basons pour la segmentation.

$$L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \quad (1)$$

$$H = \begin{cases} 256\frac{L+1}{R+1} & \text{si } R > L, \\ 255 & \text{sinon ou si } L < \alpha. \end{cases} \quad (2)$$

Le seuil α utilisé dans la transformée couleur permet de mettre en avant les zones très sombres comme l'intérieur de la bouche ou les narines. Cela a pour but de renforcer les différents gradients de teinte utilisés par la suite.

Afin de séparer les lèvres du visage et étant donné que la différence entre la peau et les lèvres est plus marquée que dans l'espace linéaire RGB, nous utilisons un algorithme de classification de la teinte du visage basé sur les k-means [5] (la classification markovienne demandant beaucoup trop de temps de calcul). Cet algorithme utilise trois classes : le visage, les lèvres et le fond.

Il est basé sur deux informations :

- le moyennage de la teinte de l'image dans un voisinage donné,
- l'écart maximum calculé entre un point de ce voisinage et la moyenne trouvée sur celui-ci.

Afin d'obtenir une bonne convergence de l'algorithme des k-means, il faut déduire automatiquement une bonne initialisation des barycentres des classes.

Pour celà de nombreux tests nous ont permis d'établir des valeurs empiriques dépendant seulement de la teinte globale du visage dans l'image (facilement calculable par histogramme), permettant une bonne initialisation de ces barycentres et donc une bonne convergence de l'algorithme vers les classes recherchée (Fig. 1). De plus nous avons établi un voisinage suffisant permettant d'obtenir des formes de lèvres lisses.

Visuellement, on sait que la bouche est la forme rouge la plus grande du visage, ainsi nous pouvons implanter un outil de recherche permettant de repérer cette zone dans la classe "lèvres". Pour cela nous parcourons les pixels appartenant à la classe "lèvres" avec un masque de filtrage non-linéaire (Eq.3) qui vérifie si le voisinage du pixel traité appartient à la même classe, et qui calcule la grandeur de chaque forme trouvée.

$$M(i) = \begin{cases} 1 + \sum_{j \in \nu(i)} a(j)M(i) & \text{si } H(i) \in \text{“}lèvres\text{”} \\ 0 & \text{sinon} \end{cases} \quad (3)$$

où M est une matrice de taille $L \times C$ (taille de l'image) initialisée à 0 et où les $a(j)$ (avec $j \in \nu(i)$) sont les coefficients attribués aux différents pixels appartenant au voisinage connexe causal du pixel traité (voir Fig. 1).

Ce procédé permet de localiser le positionnement des lèvres dans l'image en une seule passe (très rapide). Comme on peut le voir sur la Fig. 1, le masque utilisé privilégie les formes allongées horizontalement (ce qui permet de faire la différence entre les oreilles et la bouche qui sont les deux plus grandes zones de teinte rouge du visage).

Cela permet d'établir une carte du visage avec les valeurs les plus élevées sur les plus grandes formes rouges, dont la bouche avec la valeur la plus élevée de la carte.

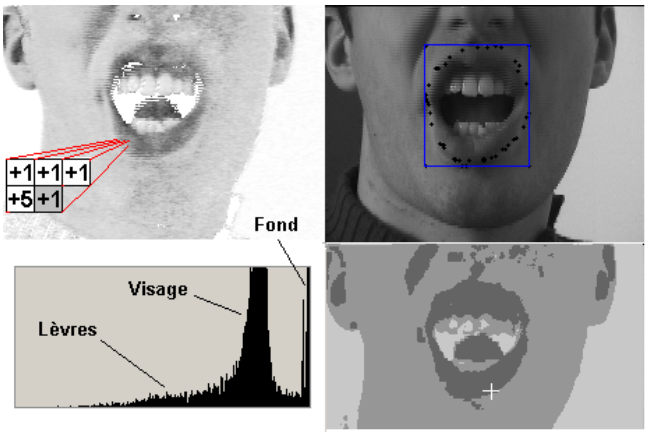

FIG. 1 – a) Teinte H et masque de filtrage. b) Suivi de points d'intérêt et cadrage de la bouche. c) Histogramme de la teinte. e) Image des classes résultant de l'algorithme k-means, point de M de plus forte valeur en blanc.

Cette méthode permet de localiser la bouche dans l'image mais pas dans sa globalité. En effet, dans certains cas, les commissures des lèvres sont peu marquées et les formes de la lèvre inférieure et de la lèvre supérieure sont séparées (p.ex. lors de la prononciation du phonème [a]). Pour contrer cela nous effectuons le suivi des formes des lèvres par l'algorithme de Lucas-Kanadé en suivant quelques points pertinents du contour externe des lèvres.

Le fait d'utiliser un algorithme d'estimation de mouvement au lieu d'une détection de mouvement [3], permet une meilleure robustesse aux conditions d'éclairage. Selon notre hypothèse, le locuteur a la bouche fermée à la première image (position au repos). Cela permet d'être sûr de bien suivre les deux lèvres durant le reste de la séquence (car la bouche fermée est détectée comme une seule forme).

Grâce à cette information supplémentaire, nous complétons la carte du visage (en recherchant une ou deux formes proches des points de suivi et en les reliant) pouvant ainsi mettre en avant la zone labiale dans sa totalité. Nous définissons ensuite

un cadre contenant de façon optimale la zone en question, par un simple parcours de la forme connexe contenant le point détecté comme étant sur les lèvres (point de M dont la valeur est la plus élevée i.e. sur la forme détectée la plus grande Fig.1.b)).

2.2 Contour actif et Animation d'avatar

Une fois le cadre contenant les lèvres défini, nous initialisons un contour actif (ou snake [6]) sur celui-ci et nous le laissons converger sur les contours externes de la bouche [4]. Le snake utilisé contient un nombre fini de points de contrôle. Ces points sont contraints à un déplacement vertical et sont initialisés sur des courbes cubiques calculées à partir de la ROI (région d'intérêt) et de la carte des lèvres (pour placer les commissures) comme présenté sur la Fig. 2.

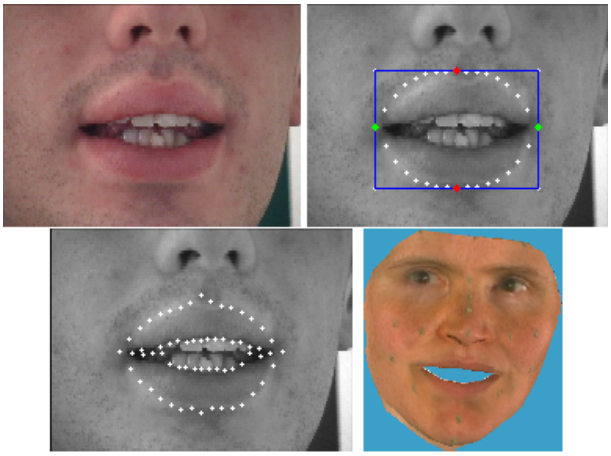

FIG. 2 – a) Entrée vidéo ; b) Initialisation du snake externe à partir de la ROI (bleu) ; c) snake interne et externe après convergence ; d) clone animé.

Les forces utilisées pour la convergence du snake sont (mise à part la force interne qui contrôle l'élasticité et la courbure de celui-ci) :

- la force externe : elle caractérise les éléments vers lesquels on veut attirer le snake sur l'image. Dans notre cas, nous utilisons un gradient calculé à partir de la teinte non-linéaire et de l'intensité lumineuse.
- la force de contrainte (spécificités du problème) : nous forçons le snake à converger vers le centre de gravité de la ROI.

Après convergence du snake externe, on initialise un second snake sur celui-ci, puis on le réduit par homothétie non isotrope par rapport au centre de la bouche (en tenant compte de l'épaisseur des lèvres traitées) et on le laisse converger sur le contour interne des lèvres.

Une fois la convergence des snakes effectuée, nous pouvons interpréter plusieurs points de contrôle de ceux-ci pour estimer les différents paramètres du clone. Ce travail est spécifique au clone utilisé. En effet, chaque interface servant à l'animation d'un avatar possède ses propres paramètres, cela nécessite donc un apprentissage différent à chaque modèle de clone. Le clone utilisé dans notre étude est un clone en 3D de 275 points mobiles fourni par l'ICP[1] qui permet d'animer de façon réaliste la bouche grâce à six paramètres d'animation sur les différents phonèmes de la langue française [7].

Nous avons donc mis en place un système linéaire qui transforme des points de contrôle des snakes en paramètres d'animation acceptables visuellement par le clone. Les points de contrôle actuellement exploités sont les suivants : la commissure droite, la commissure gauche, la position de la lèvre supérieure et la position de la lèvre inférieure.

[1] Institut de la Communication Parlé, Grenoble - http ://www.icp.inpg.fr

3 Résultats

Nous présentons deux types de résultat : les résultats d'analyse présentant la segmentation des lèvres dans l'image, la création d'une ROI, et l'extraction des contours labiaux. Puis les résultats de synthèse présentant l'animation du clone à partir de l'interprétation des points trouvés précédemment.

3.1 Résultat de segmentation (analyse)

L'espace couleur utilisé permet de réaliser une bonne segmentation même dans des conditions d'éclairage défavorable. On peut voir sur la Fig. 3 que la teinte utilisée sépare bien les lèvres du reste du visage dont la teinte reste assez homogène même avec un fort contraste ou avec un éclairage non homogène (typ. de côté).

FIG. 3 – a) Image Originale avec barbe et éclairage latéral ; b) Teinte utilisée ; c) Teinte de l'espace couleur HSL ; d) Teinte G/R souvent utilisée en segmentation labiale.

L'algorithme est robuste aux conditions d'éclairage : les changements brutaux d'éclairage (typ. ajout et suppression d'une source lumineuse) n'influent pas sur celui-ci car les différentes teintes sont estimées à chaque image à partir de la teinte moyenne du visage dans l'image courante. L'algorithme étant basé sur la teinte des lèvres sans connaissance a priori du visage (mise à part la teinte moyenne estimée de celui-ci), les visages avec barbe ou moustaches ne constituent pas un obstacle à la détection. La segmentation labiale proposée permet de localiser précisemment la bouche dans le visage, la ROI résultant de la segmentation englobe entièrement les lèvres, ce qui permet une bonne initialisation des snakes et ainsi une bonne convergence.

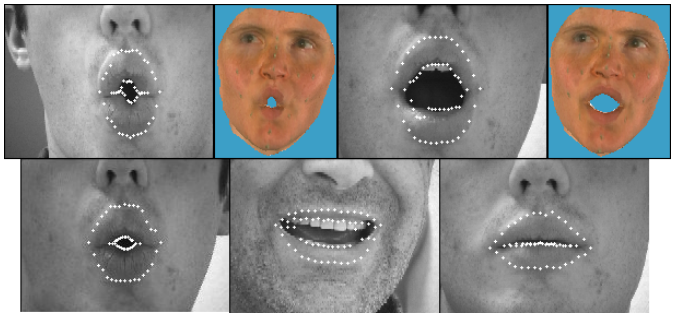

FIG. 4 – Exemples avec différents phonèmes : $[o]$ + clone, $[a]$ + clone, $[e]$, $[\partial]$ et enfin la position neutre.

3.2 Résultat d'animation (synthèse)

Une fois la ROI détectée et la convergence des contours actifs effectuée (cf. Fig.2), nous pouvons interpréter les contours détectés pour l'animation labiale du clone. Même si les contours ne sont pas forcément bien estimés, pour une animation réaliste, une bonne exploitation des résultats est plus pertinente qu'une bonne segmentation.

Dans le cas du contour interne, les dents ne posent aucun problème (elles amplifient même le gradient assurant ainsi une meilleure convergence). Le seul problème restant est l'apparition de la langue car elle possède la même teinte que la bouche. On peut remarquer que les configurations de bouche ouverte sont favorables aux dessins des contours externes. Au contraire, les configurations avec protrusion des lèvres ou bouche fermée sont favorables à l'estimation des contours internes. Nous pouvons donc coupler ces informations afin de réaliser une animation réaliste en pondérant ces valeurs selon la configuration de la bouche. Ainsi avec quelques points obtenus par une segmentation correcte et rapide, tout en respectant les proportions de la bouche, l'animation temps-réel résultante permet la compréhension des voyelles prononcées, le mouvement du clone étant très proche des mouvements réalisés par le locuteur.

4 Conclusion et Perspectives

Ce travail a permis de mettre en place un ensemble de méthodes rapides exploitant l'information fournie par une teinte non-linéaire robuste à l'éclairage, afin d'extraire les contours interne et externe des lèvres puis de les interpréter afin d'animer de façon réaliste un clone de synthèse tout en tenant compte de l'aspect temps-réel. Cela nous a permis de réaliser une chaîne complète Analyse/Synthèse temps-réel pour l'animation d'un clone sans exploitation d'information sonore et sans aucune base d'apprentissage.

L'algorithme complet d'analyse s'exécute en temps-réel, notre implantation en code C (non optimisé) permet d'atteindre une cadence vidéo supérieur à $30Hz$ avec un processeur de type $i386$ de $1.4Ghz$, soit 30 fois plus rapide que l'algorithme initial proposé dans [3].

Nos travaux actuels visent à réaliser la segmentation de l'ensemble du visage nous permettant d'obtenir des points d'intérêt précis et de pouvoir animer un modèle 3D de visage qui nous permettra de réguler les données récoltées grâce à rigidité du modèle (cf. Fig. 5).

De plus le fait d'avoir des informations sur le reste du visage nous permettra une meilleure compréhension des phonèmes prononcés (par exemple, pour faire la différence entre le phonème [e] et le phonème [y], la position du nez est très importante pour l'animation, car visuellement la bouche a pratiquement la même forme, sauf que pour le phonème [y], elle est nettement plus proche du nez que pour le phonème [e]).

Enfin nous cherchons à permettre l'animation d'un plus grand nombre de clones par l'utilisation de la norme MPEG4.

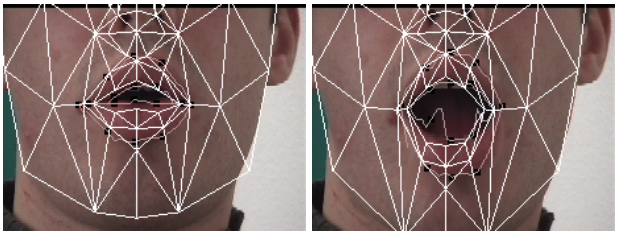

FIG. 5 – Régularisation de la forme labiale par un modèle 3D, on peut voir sur l'image de droite, que le snake interne a mal convergé à cause de la langue, le modèle 3D permet de corriger le problème.

Références

[1] T. Chen et R.R. Rao. *Audio-Visual Integration in Multimodal Communication*. Proceedings of the IEEE, Vol. 86, pp. 837-852, May 1998.

[2] T.F. Cootes et C.J. Taylor *Statistical Models of Apparence for Computer Vision* Imaging Science and Biomediacal Engineering, University of Manchester

[3] M. Liévin et F. Luthon. *Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video*. IEEE Trans. on Image Processing, Vol. 13, No. 1, pp. 63-71, Jan. 2004.

[4] P. Delmas *Extraction des contours de lèvres d'un visage parlant par contours actifs, Application à la communication multimodale* PhD Thesis, National Polytechnic Institute, Grenoble, France, Apr. 2000.

[5] J. B. MacQueen. *Some Methods for classification and Analysis of Multivariate Observations* Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, pp. 281-297, 1967.

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

[7] C. Benoît, T. Lallouache, T. Mohamadi et C. Abry. *A set of French visemes for visual speech synthesis* Talking Machines : Theories, Models, and Desgins, Elsevier Science Publishers, pp. 485-503, 1992.