Un modèle déformable paramétrique pour la reconnaissance de visages et le suivi du mouvement des lèvres

Bertrand LEROY et Isabelle L. HERLIN

Projet AIR, INRIA - Rocquencourt - BP 105 - 78153 Le Chesnay cedex - France

Cet article a pour but de présenter une méthode de localisation et d'extraction de différents éléments du visage afin de participer à l'identification de la personne et d'effectuer une mesure de l'activité labiale permettant une compréhension visuelle de la parole en environnement bruité. La localisation des régions contenant des éléments significatifs du visage, tels que la bouche, les yeux et le nez, est effectuée à l'aide d'opérateurs de morphologie mathématique. Afin d'extraire ensuite le contour de ces structures, nous employons une méthode de contour actif dédiée. Cette méthode est fondée sur l'utilisation de descripteurs de Fourier, ce qui permet d'adapter simplement le modèle à la morphologie de chacun des objets du visage.

In this paper we present a method to locate and extract salient face features in order to allow face identification and lips reading in noisy environnement. The localization of the salient features, such as nose, mouth and eyes, is obtained by using morphological operators. In order to extract the boundary of these elements, we use a specific active contour method. This model is based on Fourier descriptors and is able to incorporate information about the global shape of each object.

1 Introduction

L'extraction des différents éléments du visage constitue un module fondamental pour les systèmes de reconnaissance de visage par approche géométrique. Ce type de méthode consiste, dans un premier temps, à extraire la position relative des différents éléments caractéristiques du visage (tels que le nez, la bouche et les yeux). Différentes méthodes statistiques, inspirées des algorithmes classiques de reconnaissance de formes, sont alors utilisées pour discriminer les visages à partir de ces mesures.

On retrouve également ce processus dans le cadre de la lecture labiale de la parole. Partant du constat que la compréhension de la parole est un phénomène bi-modal, utilisant non seulement le signal auditif mais également le signal visuel, la lecture labiale se propose de décrypter la partie visuelle de la parole. Un système automatique de lecture labiale contient donc nécessairement un module permettant de connaître, à différents instants, la position et la forme des lèvres.

Dans la section 2, nous présentons donc une méthode permettant de localiser les éléments significatifs du visage tels que les yeux, le nez et la bouche. La section 3 propose ensuite un modèle spécifique de contour actif à base de descripteurs de Fourier afin d'extraire la frontière de ces éléments caractéristiques. Ce modèle paramétrique offre l'avantage de pouvoir s'adapter aisément à la géométrie de l'objet étudié. Les résultats, concernant l'extraction des éléments du visage ou le suivi du mouvement des lèvres, sont présentés dans la section 4.

Cette étude a été effectuée dans le cadre du projet AMIBE (projet de recherche subventionné par le PRC homme-machine), qui se propose d'expérimenter une interface homme-machine intégrant le son et l'image dans le cadre d'une transaction de guichet bancaire. L'utilisateur est identifié simultanément par sa voix et son visage. Les données utilisées sont donc des séquences audio-vidéo de locuteurs filmés de face avec un éclairage fixe.

2 Localisation d'éléments du visage

L'utilisation de modèles de contour actif afin d'extraire les éléments du visage tels que la bouche, le nez et les yeux, suppose leur localisation préalable. Ces éléments sont des structures peu mobiles mais elles peuvent subir des déformations. Elles représentent généralement un contraste important par rapport à leur environnement. Le processus de localisation est fondé sur l'utilisation des particularités de l'acquisition.

La première tâche définie consiste à localiser les yeux. Le nez et la bouche seront ensuite recherchés en utilisant la géométrie du visage : ils intersectent généralement la médiatrice du segment défini par les deux yeux et leur localisation exacte est donc obtenue en effectuant une recherche le long de cet axe. L'existence d'une tache de réflexion spéculaire, due aux sources lumineuses frontales sur l'iris, nous permet de localiser l'œil en utilisant un opérateur morphologique de type "pic", qui a pour effet de faire ressortir les maxima locaux de la fonction d'intensité de niveaux de gris. Après application de cet opérateur à l'image de niveaux de gris, la position

des iris correspond aux deux régions où l'intensité sur l'image résultat est la plus importante.

Pour localiser le nez et la bouche, nous utilisons l'information spatio-temporelle de façon à caractériser les points ayant enregistré les plus fortes variations d'intensité au cours de la séquence. À cet effet, nous réalisons en chaque point la sommation de la norme du gradient spatio-temporel (voir figure 1). Cette image I_{som} est obtenue par :

$$I_{som} = \sum_{i=1}^{N} \sqrt{(\frac{\partial I}{\partial x})^2 + (\frac{\partial I}{\partial y})^2 + (\frac{\partial I}{\partial t})^2}, \qquad (1)$$

où $I(x, y, t)$ représente l'intensité de niveau de gris ; x, y les composantes spatiales ; t la composante temporelle et N le nombre d'images de la séquence.

Sur l'image I_{som}, on définit des régions de forte intensité et parmi elles les régions de la bouche et du nez, qui sont obtenues par intersection avec la médiatrice du segment passant par les iris.

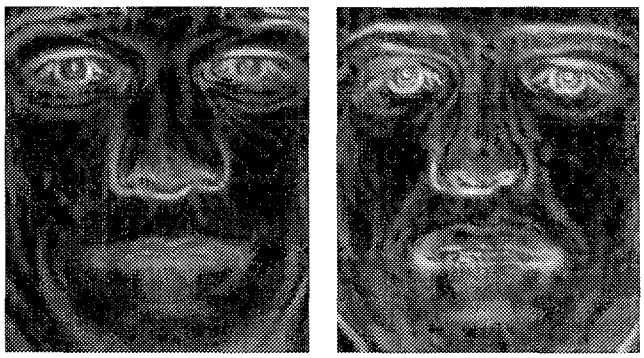

FIG. 1 - *Calcul de la norme du gradient spatio-temporel sur deux séquences différentes.*

Afin d'accélérer le traitement informatique, les images utilisées pour la localisation sont les trente premières de la séquence vidéo, ce qui correspond à une durée d'acquisition proche de la seconde. Utiliser une séquence plus courte aurait pour effet de diminuer la précision des résultats, car la norme de la composante temporelle du gradient est assez faible.

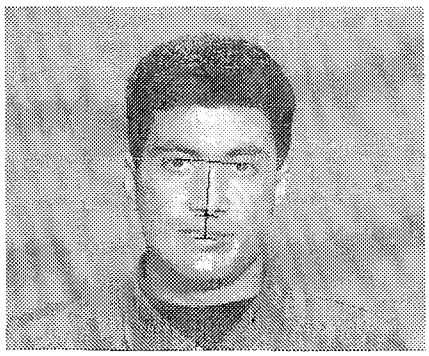

FIG. 2 - *Résultat de la localisation des yeux, du nez et de la bouche.*

3 Extraction d'éléments du visage

L'utilisation d'un modèle de contour actif, afin d'extraire les éléments du visage, suppose que l'on dispose d'une méthode capable de détecter avec précision la frontière de structures pouvant subir de fortes déformations (pour permettre la caractérisation spatio-temporelle de la bouche) et possédant des coins (pour permettre l'étude des yeux et de la bouche). Les modèles de contour actif génériques tels que les "snakes" [KWT87] [CC90] ne permettent pas de prendre en compte la forme spécifique des objets. Plus exactement, les contraintes géométriques appliquées à ces modèles apparaissent sous la forme de paramètres énergétiques et ne sont pas intégrées au modèle.

Pour toutes ces raisons, des études spécifiques ont été réalisées afin de définir des modèles adaptés à l'extraction des contours d'éléments du visage ; ces modèles sont généralement des modèles déformables paramétriques. Yuille, Cohen et Hallinan [YCH88] ont ainsi proposé un ensemble de modèles déformables spécifiques à chaque élément du visage.

Craw, Tock et Bennett [CTB92] ont décrit un modèle déformable global permettant d'extraire près de 40 points caractéristiques sur une image de visage. Ce modèle est constitué d'un ensemble de modules adaptés au traitement de chacun des éléments spécifiques du visage et d'une structure globale contrôlant la position des différents modules les uns par rapport aux autres. Chaque module est gouverné par un ensemble de contraintes statistiques qui lui est spécifique. Cette méthode donne des résultats intéressants concernant la localisation des éléments du visage ; cependant, elle ne permet pas d'extraire avec précision les contours de ces éléments et nécessite un temps de calcul important.

Nous présentons donc dans cet article un modèle paramétrique de contour actif à base de descripteurs de Fourier. Notre objectif est de disposer d'un modèle simple et suffisamment général pour pouvoir extraire la bouche, les yeux et le nez tout en tenant compte de la géométrie propre à chacun de ces objets. L'utilisation de descripteurs de Fourier est particulièrement adéquate dans le cas de la bouche et des yeux puisque la forme de ces objets est proche de celle d'une ellipse.

3.1 Courbes elliptiques fermées

Une courbe elliptique fermée peut-être représentée mathématiquement sous la forme suivante :

$$v_N(\theta) = \begin{pmatrix} x_N(\theta) \\ y_N(\theta) \end{pmatrix} = \sum_{k=0}^{N} A_k \begin{pmatrix} \cos(\theta k) & \sin(\theta k) \\ \cos(\theta k) & \sin(\theta k) \end{pmatrix}, \qquad (2)$$

où A_k est une matrice 2×2 et N le nombre d'harmoniques utilisées pour décrire cette courbe.

Staib et Duncan [SD92] ont été les premiers à proposer l'utilisation de tels descripteurs pour modéliser des contours actifs. Leur méthode de convergence est une méthode stochastique bayesienne. Nous proposons un modèle de contour actif à base de descripteurs de Fourier utilisant une méthode variationnelle.

Afin d'extraire les contours d'un objet, nous recherchons un minimum local de la fonctionnelle d'énergie :

$$E(v) = \int_0^{2\pi} P(v(\theta)) + \lambda(v_\theta(\theta))^2 d\theta, \text{ où } P = -|\nabla I * G|^2 \text{ et } \lambda \in \Re^+$$

Le premier terme de l'énergie est lié à l'image traitée. P correspond au carré de la norme du gradient calculé dans l'image I convoluée par un filtre de lissage G. Le second terme de l'énergie est un terme d'élasticité car $v_\theta^2(\theta)$ est une mesure de l'énergie liée à la tension de la courbe. Le paramètre λ est un poids associé à la contrainte d'élasticité et il permet d'effectuer une pondération entre les forces d'élasticité et les forces issues du potentiel image.

La recherche de la courbe v^*, réalisant un minimum local de E, est effectuée à l'aide d'une méthode de minimisation de type Newton. Afin d'assurer une meilleure convergence, la minimisation est d'abord effectuée sur les paramètres de basses fréquences puis sur les paramètres de fréquences plus élevées.

3.2 Courbes elliptiques ouvertes connexes

L'utilisation de courbes elliptiques fermées est problématique lorsque le contour de l'objet recherché possède des points fortement irréguliers (tel que le contour des yeux) au sein de portions régulières. Ainsi, pour pouvoir modéliser des coins, il est nécessaire d'utiliser des paramètres de fréquences élevées, ce qui a pour effet de créer des oscillations sur les portions de courbes devant être régulières. Afin de remédier à ce problème, nous utilisons des modèles elliptiques ouverts. Les courbes ouvertes sont généralement utilisées pour représenter des contours d'objet n'ayant pas de frontière complète ou pour décrire des segments de courbe comme une lettre [PF86]. Dans le cas présent, un modèle formé de deux courbes elliptiques ouvertes et connexes permet de représenter de façon fiable le contour de la bouche ou des yeux ; celui du nez est par contre décrit par une seule courbe ouverte.

Soit v_N un modèle formé de deux courbes ouvertes jointes en leurs extrémités v_A et v_B et défini par :

$$v_N(\theta) = \begin{cases} v_A(\theta) = \begin{pmatrix} x_a(\theta) \\ y_a(\theta) \end{pmatrix} & \text{si } 0 \leq \theta < \pi \\ v_B(\theta) = \begin{pmatrix} x_b(\theta) \\ y_b(\theta) \end{pmatrix} & \text{si } \pi \leq \theta < 2\pi \end{cases} \quad (3)$$

v_A et v_B respectent de plus les conditions :

$$v_A(0) = v_B(2\pi) \text{ et } v_A(\pi) = v_B(\pi). \quad (4)$$

La fonctionnelle d'énergie associée au modèle devient :

$$E(v_N) = \int_0^\pi P(v_A(\theta)) + \lambda(\frac{\partial v_A(\theta)}{\partial \theta})^2 d\theta + \int_\pi^{2\pi} P(v_B(\theta)) + \lambda(\frac{\partial v_B(\theta)}{\partial \theta})^2 d\theta. \quad (5)$$

L'utilisation de descripteurs de Fourier dans le cas de courbes ouvertes nécessite la définition d'un nouveau modèle. En effet, afin d'être cohérent avec la formulation de l'énergie définie par l'équation (5), il convient de choisir une base différente de celle utilisée en (2). Nous choisissons d'utiliser une base de cosinus pour décrire $x(\theta)$ et de sinus pour décrire $y(\theta)$. Ces bases sont orthogonales sur $[0, \pi]$ et il suffit d'une seule harmonique pour décrire une demi-ellipse lorsque θ varie de 0 à π. L'ensemble des courbes décrites par ces bases est représenté par :

$$v_A(\theta) = \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} + R_\phi \sum_{k=1}^{N} \begin{pmatrix} a_k \cos(\theta k) \\ b_k \sin(\theta k) \end{pmatrix}, \quad (6)$$

où a_k et b_k sont les paramètres du modèle, N le nombre d'harmoniques utilisées pour décrire la courbe et R_ϕ la matrice de rotation d'angle ϕ par rapport à l'origine. Le déphasage existant entre les descripteurs utilisés pour $x(\theta)$ et $y(\theta)$ permet de décrire une demi-ellipse sans être obligé d'utiliser un nombre important d'harmoniques.

Dans la pratique, les contours de la bouche sont modélisés par une courbe à une harmonique pour la partie inférieure et cinq pour la partie supérieure. De même, on utilise une harmonique pour la partie inférieure des yeux et trois pour la partie supérieure.

Soit v_A une courbe définie dans une base elliptique de Fourier et décrite par l'équation (6), la courbe v_B décrite par une seule harmonique et répondant aux conditions données en (4) est représentée par :

$$v_B(\theta) = \begin{pmatrix} \sum_{i=0}^{N/2} a_{2i} \\ b_0 \end{pmatrix} + R_\phi \begin{pmatrix} \left(\sum_{i=0}^{N/2} a_{2i+1}\right) \cos(\theta) \\ c \sin(\theta) \end{pmatrix}. \quad (7)$$

Un examen rapide de l'énergie $E(v)$ montre que les dérivées partielles par rapport aux paramètres a_k n'ont pas le même poids que celles calculées par rapport aux paramètres b_k et c. Ainsi, les forces exercées sur l'axe horizontal sont deux fois supérieures aux forces verticales, toute chose égale par ailleurs. Pour améliorer la convergence, nous pondérons donc les gradients de $E(v)$ par rapport aux paramètres a_k par un facteur de 0.5. Les résultats obtenus par une minimisation de type Newton, en utilisant cette pondération, sont des minima locaux de $E(v)$; il n'est pas nécessaire de modifier la fonctionnelle d'énergie.

Si l'on se place dans le cas particulier d'un potentiel image nul, le résultat obtenu en minimisant la fonctionnelle d'énergie à partir d'une courbe initiale représentée par une ellipse est identique avec ou sans pondération : c'est une courbe de longueur nulle. Cependant, les courbes obtenues à chaque itération de la descente du gradient sont différentes. Sans pondération, le grand axe se réduit plus rapidement que le petit axe et la courbe devient un cercle. Avec une pondération, le rapport du grand axe au petit axe reste constant. Ainsi, la force de rétraction appliquée à la courbe n'a pas de caractère régularisant. Il est alors possible d'influencer la géométrie de la courbe obtenue au minimum local de la fonctionnelle d'énergie par le biais de contrainte sur l'initialisation.

4 Résultats

Les algorithmes utilisés pour détecter les contours des éléments du visage sont de type pyramidal. Ainsi, partant d'une ellipse comme courbe initiale entourant l'objet étudié, on recherche une courbe v_1 décrite par une seule harmonique. Le résultat obtenu est alors utilisé comme initialisation d'un modèle v_2, à deux harmoniques, jusqu'à l'obtention d'une courbe v_N, à N harmoniques, décrivant les contours de l'élément recherché.

La force d'élasticité qui pousse le contour à se rétracter vers l'objet joue un rôle déterminant lors de la première étape. Dans les étapes suivantes, le nombre d'harmoniques augmente et cette force exerce uniquement un effet de rappel sur la courbe. Cette méthode permet de converger, rapidement et sûrement, vers une approximation du contour de l'objet.

4.1 Obtention des différents éléments du visage

Les figures 3 et 4 donnent un aperçu des résultats obtenus à différents moments du processus de recherche des contours du nez et de la bouche. L'expérience montre qu'une courbe elliptique ouverte à cinq harmoniques permet de modéliser les contours du nez avec une précision satisfaisante. La bouche est décrite par un modèle de deux courbes ouvertes connexes utilisant respectivement cinq harmoniques pour la partie supérieure et une harmonique pour la partie inférieure.

FIG. 3 - *Extraction du contour du nez. Initialisation et résultats obtenus avec une, puis cinq harmoniques.*

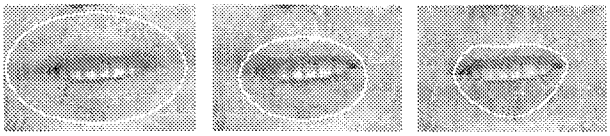

FIG. 4 - *Extraction du contour de la bouche. Initialisation et résultats obtenus avec une, puis cinq harmoniques.*

Le voisinage des yeux forme une région riche en contours (l'iris, les paupières, les sourcils, ...) et il est nécessaire d'utiliser une initialisation précise afin de détecter les yeux. Cette détection est effectuée en deux temps. Dans un premier temps, la tache de reflexion spéculaire située à l'intérieur de l'iris est localisée (voir section 2) et on recherche les contours de l'iris. Le diamètre et la position de l'iris nous permettent alors d'initialiser le modèle de contour actif de sorte que la courbe ne soit pas attirée par des contours externes aux yeux. La figure 5 montre les résultats obtenus à l'issue des différentes étapes de l'algorithme.

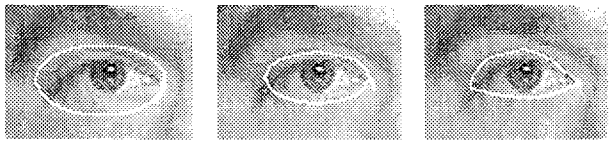

FIG. 5 - *Extraction du contour des yeux. Initialisation et résultats obtenus avec une, puis trois harmoniques.*

4.2 Suivi du mouvement labial

Les systèmes de lecture labiale existants, incorporent généralement des modules permettant l'acquisition des données labiales à partir de marqueurs préalablement placés sur les lèvres du locuteur. Ces marqueurs sont soit des pastilles de couleur vive placées sur les lèvres, soit un maquillage des lèvres.

Pentland & Mase (voir [PM89]) ont proposé un système de lecture labiale fondé sur le calcul du flot optique sans qu'il soit nécessaire d'utiliser des marqueurs. Nous proposons une approche moins coûteuse en calculs et utilisant la méthode de détection de contour à base de courbes elliptiques, de façon à permettre un lien entre l'identification du visage et la lecture labiale.

Le suivi du mouvement des lèvres est obtenu en minimisant la fonctionnelle d'énergie (5) pour chaque image de la séquence : la courbe résultat sur une image sert d'initialisation sur l'image suivante. Nous avons ainsi pu obtenir des informations pertinentes pour l'analyse de la parole telles que l'étirement, l'ouverture, la surface et les dérivées temporelles de ces valeurs.

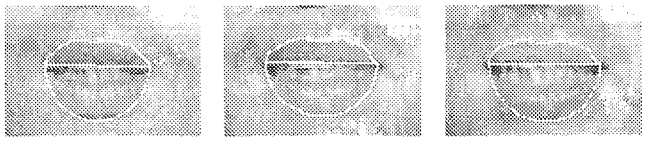

FIG. 6 - *Suivi du mouvement des lèvres.*

Références

[CC90] Laurent D. Cohen and Isaac Cohen. A finite element method applied to new active contour models and 3-D reconstruction from cross sections. In *IEEE Proceedings of the Third International Conference on Computer Vision*, pages 587–591, Osaka, Japan, December 1990.

[CTB92] I. Craw, D. Tock, and A. Bennett. Finding face features. In *ECCV'92*, pages 92–96, 1992.

[KWT87] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. In *Proceedings of the First International Conference on Computer Vision*, pages 259–268, London, June 1987.

[PF86] E. Person and K. Fu. Shape discrimination using fourier descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):388–397, May 1986.

[PM89] A. Pentland and K. Mase. Automatic visual recognition of spoken words. Technical Report 117, M.I.T. Media Lab Vision Science Technical Report, 1989.

[SD92] L.H. Staib and J.S. Duncan. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, November 1992.

[YCH88] A. Yuille, D. Cohen, and P. Hallinan. Facial feature extraction by deformable templates. Technical Report 88-2, Harvard Robotics Laboratory, 1988.