Un système de détection de visage et d'extraction de paramètres basé sur les SVM et des contraintes

W. Karam[1] C. Mokbel[1] H. Greige[1] B. Pesquet-Popescu[2] G. Chollet[2]

[1] Computer Science Department, University of Balamand, PO Box 100 Tripoli, Lebanon
[2] Ecole Nationale Supérieure des Télécommunications, 46 rue Barrault, 75634 Paris cedex 13, France

{karam, chafic.mokbel, hanna.greige}@balamand.edu.lb
{karam, pesquet, chollet}@tsi.enst.fr

Résumé

La détection de visage et de composantes faciales a suscité récemment un intérêt grandissant vu la multitude d'applications qui en découlent. Dans ce travail, la base de données de visages parlants BIOMET qui est développée à l'ENST est utilisée pour l'extraction et le suivi des composantes des visages dans les séquences vidéo. Pour ce faire, une machine SVM est apprise sur des fenêtres après leur transformation dans le domaine d'ondelettes. Un modèle géométrique statistique est ensuite appliqué afin de lisser la sortie de la machine SVM et d'affiner la détection. Un autre modèle probabiliste sur les distances aux frontières SVM permet plus de lissage et une meilleure sélection des composantes faciales. Les résultats expérimentaux montrent une bonne détection qui peut atteindre les 95% selon le protocole de mesure.

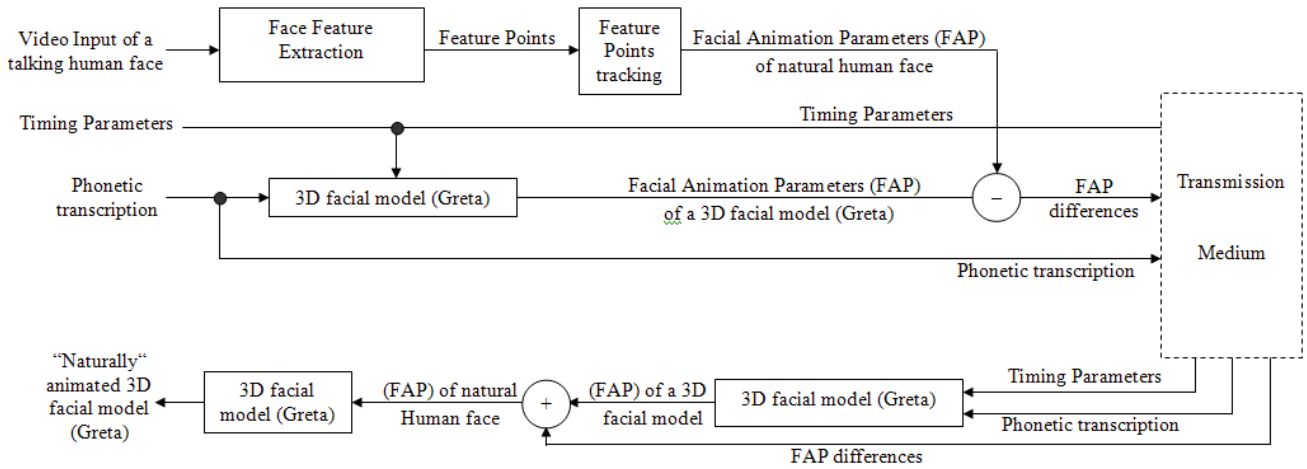
Mots clefs

Détection de visage, Composantes faciales, SVM, Reconnaissance statistiques des formes, Visages parlants.

1 Introduction

La détection automatique de visage et de paramètres descriptifs dans le visage est un sujet de grand intérêt. On trouve des applications dans divers domaines et plus particulièrement l'indexation et la recherche automatique de documents visuels ou audiovisuels, et le codage par indexation de documents. Le travail décrit dans ce papier s'insère dans le domaine du codage par indexation de scènes video. Le but est d'extraire, au codage, des paramètres descriptifs du visage afin d'animer au décodage un modèle de synthèse de visages parlants 3D, Greta [1]. Le système final est bimodal et on se sert de la parole pour aider dans cette tâche de transcription. L'ensemble du système final est décrit dans la Figure 1. Dans ce travail nous nous intéressons au premier module qui consiste à détecter et à suivre quelques paramètres spéciaux du visage, tels que les yeux, les lèvres et le nez, dans une scène audiovisuelle. Pour ce faire, un système à base de machines à vecteurs de support (''Support Vector Machines'' SVM) a été appris et utilisé pour détecter les paramètres d'animation faciale (''Facial Animation Parameters'' FAP). Dans [2] l'utilisation des SVMs dans cette tâche de détection donne des résultats satisfaisants. De plus, il est montré que la détection de composantes du visage avec quelques contraintes donne de meilleurs résultats que la détection direct du visage.

Dans ce papier, nous décrivons un système, basé sur les SVMs, de détection de visages en passant par la détection



de composantes du visage. La machine de détection est paramétrée de manière à surdétecter et les résultats en sortie sont lissés par différentes contraintes. D'abord un modèle probabiliste sur la distance de la composante détectée au vecteur de support le plus proche est utilisé. Ensuite et sur le plan géométrique, un modèle statistique sur les composantes du visage est construit et appliqué trame par trame. Des contraintes sur le déplacement temporel des composantes détectées sont aussi proposées.

Dans la section suivante, les techniques de détection de visage sont brièvement revues. La section 3 décrit les SVMs et leur utilisation en détection de composantes faciales. La section 4 décrit le modèle probabiliste que nous introduisons sur la sortie de la machine SVM. La section 5, un modèle statistique de contraintes géométriques est présenté. Les expériences et résultats sont donnés à la section 6. Finalement, le papier se termine par des conclusions et des perspectives.

2 Détection de visage

La détection de visages et des composantes faciales a connu un grand intérêt récemment vu l'utilité dans diverses applications telles que l'indexation, la gestion du contenu, la reconnaissance de la parole audiovisuelle [3].

Une revue des techniques de détection de visages est effectuée dans [4][5]. Yang et. al. [3] classifient ces techniques en 4 classes : techniques descriptives basées sur la connaissance, techniques basées sur l'extraction de paramètres cractéristiques invariants, techniques basées sur la superposition de caractéristiques, et techniques basées sur l'apparence. Dans la dernière classe, les techniques utilisent l'analyse statistique et l'apprentissage automatique pour construire des machines capables de séparer les visages des non-visages. Les réseaux de neurones, les machines à vecteurs de support (SVM), les classifieurs bayesiens, les modèles de Markov cachés (HMM) sont parmi les techniques d'apprentissage automatique les plus souvent utilisées.

Dans ce papier, un système à base de SVM est présenté. Les vecteurs de support à fonction polynomiales sont appris sur des images extraites de la base de données BIOMET[1].

3 Machines à vecteurs de support

La théorie d'apprentissage statistique (''*Statistical Learning Theory* SLT) a été présentée par Vapnik [6]. Dans ce cadre, un algorithme d'apprentissage est un algorithme qui détermine automatiquement la meilleure fonction qui décrit la relation entre l'entrée et la sortie d'une machine en se basant sur un nombre limité d'exemples. Cette technique permet de trouver une surface qui sépare au mieux les classes de données en maximizant la marge entre ces classes. A la différence des approches d'apprentissage se basant sur la minimization du risque empirique, risque mesuré sur les données d'apprentissage tel que l'erreur quadratique moyenne, cette approche est basée sur le principe de *minimization du risque structuré*. Il s'agit de minimiser le majorant de l'erreur réelle. Ceci offre une capacité inhérente de généralisation de la machine trouvée [7].

Afin d'illustrer le fonctionnement des SVMs nous reprenons ici l'exemple classique des classifieurs linéaires appliqués au problème de séparation de deux classes. Supposons qu'on possède un ensemble de données $\{(x_i, y_i)\}_{i=1}^{l}$, où $y_i \in \{-1,+1\}$ et $\mathbf{x_i}$ les entrées. On cherche le classifieur linéaire qui sépare ces données avec la plus faible erreur de généralisation. Intuitivement ce classifieur est un hyperplan qui maximise la marge d'erreur, qui est la somme des distances entre l'hyperplan et les exemples positifs et négatifs les plus proches de cet hyperplan. (Figure 2).

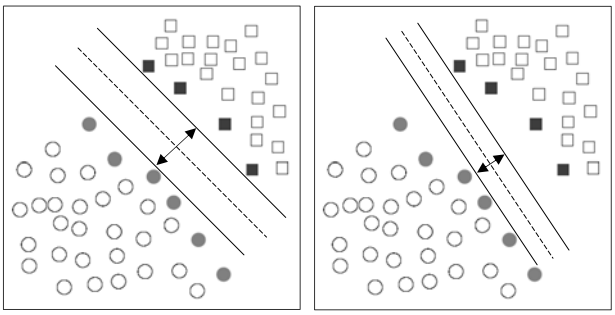


Figure 2: *Le problème de séparation linéaire entre 2 classes et les vecteurs de support (points remplis). L'hyperplan de séparation avec une marge maximale (à gauche) et une marge faible (à droite).*

Dans le cas où les données ne peuvent pas être séparées par une fonction linéaire, une non-linéarité peut être introduite grâce à l'utilisation d'une fonction symmétrique positive appelée fonction de noyau $\boldsymbol{K}$ (Figure 3). Cette fonction permet de définir une solution de la form: $f(x) = \sum_{i=1}^{N_{SV}} \alpha_i y_i K(s_i, x) + b$ où $\{\alpha_i\}$ sont des facteurs multiplicatifs positifs de Lagrange et $\boldsymbol{s_i}$ les vecteurs de support.

Dans ce papier nous utilisons les SVMs comme machines de classification qui permettent la détection de composantes du visage, à savoir les yeux, le nez et la bouche. Pour chaque composante, une machine SVM est construite et est appliquée sur des fenêtres balayant successivement une image. Une transformée en ondelette est appliquée sur chaque fenêtre avant de la traiter par la machine de classification. La machine permet de détecter si la fenêtre en entrée correspond à la composante

[1] BIOMET est une base de données de visages parlants construite à l'ENST

recherchée. Dans ce travail nous avons utilisé des machines à fonction de noyau polynomiale de degré 3. Il est évident que chaque machine SVM détectera plus d'une composante dans une image. D'où la nécessité de contraindre le système afin de sélectionner les solutions optimales. Plusieurs modèles de contraintes sont décrits dans ce qui suit.

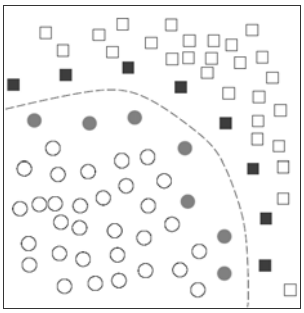


Figure 3: *Une surface de décision construite par un classifieur polynomial. Cette figure illustre le cas de données séparables par fonction non linéaire.*

4 Modèle probabiliste sur la sortie du SVM

La machine SVM permet de sélectionner la classe de la donnée en entrée en regardant le signe de la fonction $f(\mathrm{x})$ dans le cas de classifieur binaire comme dans cette étude. Il est clair que plus la valeur de $|f(\mathrm{x})|$ est grande plus le vecteur en entrée est loin de la frontière. D'où l'idée d'appliquer une méthode de sélection de la meilleure composante détectée en fonction de la valeur de $f(\mathrm{x})$ associée. Pour ce faire, on utilise un modèle gaussien centré sur l'inverse de $f(\mathrm{x})$ dont la variance est déterminée par apprentissage. La composante détectée la plus vraisemblable sera ainsi sélectionnée.

5 Modèle géométrique statistique

L'introduction de contraintes géométriques permet aussi d'améliorer la sélection des meilleures composantes parmi celles détectées par les machines SVM. L'idée est de construire un modèle statistique sur la position des composantes d'un visage par rapport à une composante centrale, à savoir le nez. Pour chaque nez détecté, les autres composantes optimales lui sont associées définissant ainsi un log-vraisemblance pour ce nez égal à la somme des log-vraisemblances des composantes associées :

$$LL_n = \sum_{c=\{re,le,m\}} LL_c = LL_{re} + LL_{le} + LL_m$$

Le nez ayant la meilleure vraisemblance sera ensuite sélectionné et avec lui l'ensemble des composantes associées. Nous décrivons dans ce qui suit le principe. Prenons, à titre d'exemple, le cas d'une composante, l'œil droit. Cette composante est caractérisée par les paramètres suivants (*Figure 4*) :

x_{re}, y_{re}: x et y sont les coordonnées par rapport au coin haut-droit de l'image.

w_{re}, h_{re}: w et h sont la largeur et la hauteur de la fenêtre glissante utilisée pour la détection de la composante. Les paramètres géométriques pour les autres composantes peuvent aussi être définis (x_{le}, y_{le}, w_{le}, h_{le}, x_m, y_m, w_m, h_m, x_n, y_n, w_n, h_n.) où *le*, *m*, et *n* indiquent œil gauche, bouche et nez respectivement.

Comme le modèle considère le nez comme référence, sans perte de généralité, il est adéquat de normaliser les coordonnées par rapport aux coordonnées du nez :

Oeil droit: $x'_{re} = x_n - x_{re}$, $y'_{re} = y_n - y_{re}$,

Oeil gauche: $x'_{le} = x_n - x_{le}$, $y'_{le} = y_n - y_{le}$,

Bouche: $x'_m = x_n - x_m$, $y'_m = y_n - y_m$,

Les coordonnées d'une composante sont supposées indépendantes et suivent une loi gaussienne :

$$\mathrm{f_{xy}}\left(\mathrm{x'_i}, \mathrm{y'_i}\right) = \frac{1}{2\pi\sigma_{\mathrm{x_i}}\sigma_{\mathrm{y_i}}} \mathrm{e}^{-\left(\frac{\mathrm{x'_i}-\mu_{\mathrm{x_i}}}{\sigma_{\mathrm{x_i}}}\right)^2 - \left(\frac{\mathrm{y'_i}-\mu_{\mathrm{y_i}}}{\sigma_{\mathrm{y_i}}}\right)^2}$$

Le vecteur moyen et la matrice de covariance supposée diagonale peuvent être estimés à partir des données d'apprentissage.

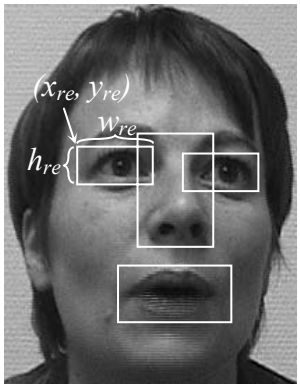


Figure 4: *Détermination manuelle des composantes et de leurs coordonnées dans les images d'apprentissage.*

6 Expériences et résultats

Les expériences sont conduites sur un sous-ensemble de la base de données Biomet définie dans l'introduction. Ce sous-ensemble consiste de 70 trames de vidéo appartenant à 7 différents visages parlants. Chaque trame est de dimensions 166 x 216. Une partie des données a été segmentée manuellement pour identifier les composantes ; œil droit, œil gauche, nez, bouche. Cette partie sera nommée ensemble d'apprentissage. L'histogramme des

niveaux de gris dans ces images a été égalisé en prétraitement [8][9] pour réduire l'effet de l'éclairage qui introduit un facteur d'échelle nuisible à la classification.

Pour chaque composante, une fenêtre de taille fixe est glissée sur toute l'image. Si la fenêtre chevauche avec plus de 70% avec la fenêtre positionnée manuellement, elle est étiquetée positivement (+1). Sinon, elle est étiquetée négativement (-1). Pour chaque fenêtre une transformation en ondelettes (Cohen-Feauveau-Daubechies 9/7) biorthogonale de 2 niveaux de décomposition est appliquée [10][11]. Les coefficients d'ondelettes de toutes les trames sont ensuite utilisés pour apprendre une machine de classification SVM avec une fonction de noyau polynomiale de degré 3. Le passage dans le domaine transformé d'ondelettes permet une représentation plus condensée des données en gardant un maximum d'information discriminante.

Une fois les machines SVM apprises, la détection des composantes du visage commence en passant des images de test dans les machines de classification SVM. Ces machines produisent généralement plus d'une détection pour chaque composante. Pour sélectionner la meilleure détection les deux approches décrites en sections 4 et 5 sont appliquées séparément ou combinées. La Figure 5 montre l'ensemble du système.

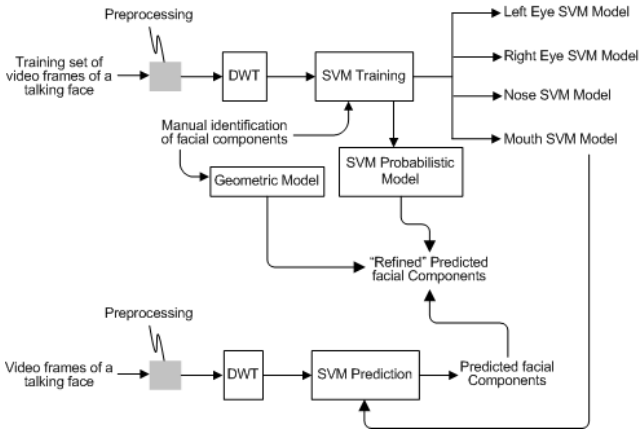


Figure 5: *Extraction des composantes du visage.*

Le tableau 1 donne les taux de bonne détection des composantes faciales. Une composante est considérée comme bien détectée si elle occupe plus de 70% de surface commune avec une détection manuelle. Il est à noter que le modèle probabiliste SVM permet une amélioration de 3.23% des performances en moyenne.

oeil droit	oeil gauche	nez	Bouche
95.71%	88.57%	88.57%	94.29%

Table 1: *Taux de détection des composantes faciales.*

7 Conclusions et perspectives

Un système de détection de visage et plus particulièrement des composantes faciales a été développé et décrit dans ce papier. Ce système est basé sur les machines à vecteurs de support utilisées comme classifieurs. Plusieurs fenêtres sont généralement détectées pour chaque composante. Dans ce papier nous avons décrit deux méthodes permettant de sélectionner les fenêtres optimales. Ces méthodes ont été combinées.

Dans la suite, nous nous intéressons à une contrainte supplémentaire par l'introduction de la dimension temporelle. Le suivi et le lissage des trajectoires temporelles des composantes faciales seront étudiés.

Par ailleurs, ce module de détection sera intégré dans le système de codage par indexation décrite dans l'introduction.

Références

[1] S. Pasquariello and C. Pelachaud: Greta: A Simple Facial Animation Engine, 6th Online World Conference on Soft Computing in Industrial Applications, Session on Soft Computing for Intelligent 3D Agents, September 2001.

[2] B. Heisele, A. Verri, and T. Poggio, "Learning and Vision Machines", *Proceedings of the IEEE*, vol. 90, no. 7, july 2002.

[3] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "*Audio-Visual Automatic Speech Recognition: An Overview*." In: Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, 2004.

[4] M-H Yang, D. J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34-58, Jan. 2002.

[5] E. Hjelmas and B. K. Low, "Face detection: A survey". *Computer Vision and Image Understanding*, 83(3), 2001.

[6] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[7] E. Osuna, R. Freund, and F. Girosi. "Support vector machines: Training and applications." Technical report A.I. Memo No. 1602, C.B.C.L. Paper No. 144, *MIT Center for Biological and Computational Learning*, 1997.

[8] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: global versus component-based approach", *Proc. 8th International Conference on Computer Vision*, vol. 2, pp. 688–694, Vancouver, 2001.

[9] B. Heisele, T. Poggio, and M. Pontil, "Face detection in still gray images," Technical Report A.I. Memo No. 2001-010, C.B.C.L. Paper No. 197, *MIT Center for Biological and Computational Learning*, 2000.

[10] A. Cohen, I. Daubechies, J. C. Feauveau, "Biorthogonal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485-560, May 1992.

[11] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, "Image Coding Using Wavelet Transform," IEEE Transactions on Image Processing, vol. 1, no. 2, pp. 205-220, April 1992.