Un modèle actif d'apparence hiérarchique pour la détection des traits de visages

Franck DAVOINE, Van Mô DANG

Laboratoire Heudiasyc, CNRS, Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex.
Franck.Davoine@hds.utc.fr, Van.Mo.Dang@hds.utc.fr

Résumé – L'article présente une solution visant à extraire les traits caractéristiques (yeux, nez et bouche) de visages. La méthode repose sur l'utilisation d'un modèle actif d'apparence calculé à partir de descriptions texturales hiérarchiques de visages d'une base d'apprentissage. Ces descriptions sont les réponses de bancs de filtres de Gabor aux niveaux de points choisis sur un visage. Le modèle d'apparence, initialement proche du visage test, épouse de manière automatique la forme du visage selon une procédure d'adaptation itérative. La qualité de l'adaptation du modèle d'apparence hiérarchique est comparée à celle du modèle plus classique d'apparence pixellique.

Abstract – *An active hierarchical appearance model to detect feature points of faces* - The article presents a solution aiming at extracting the feature points (eyes, nose and mouth) of faces. The method makes use of an active appearance model based on hierarchical descriptions of a training set of faces. These descriptions are the responses of Gabor filter bunches to sets of points chosen on a face. The appearance model, initially near to the test face, automatically adapts to the shape of the face according to an iterative adaptation procedure. The quality of the adaptation of the hierarchical appearance model is compared with that of the more traditional pixel-based model of appearance.

1 Introduction

L'analyse de visages par traitement d'images est encore aujourd'hui un domaine de recherche très actif, qui concerne de nombreuses applications telles que les interfaces homme machine, la communication au travers de réseaux à très bas débits ou les systèmes d'identification pour le contrôle d'accès. Les recherches englobent la détection, le suivi, le codage, la reconnaissance, la synthèse de visages et de leurs principaux attributs (pose, regard, lèvres, expressions, mouvements et comportement, âge, genre, occlusions, etc.). Parmi les méthodes proposées, nombreuses sont celles qui utilisent des modèles d'apparence permettant une coopération entre l'analyse et la synthèse d'un visage [10, 4, 6, 5]. Dans cet article, nous nous intéressons plus particulièrement au modèle actif statistique d'apparence *(AAM)*, initialement proposé par Cootes et al. [2], et qui permet de contrôler à la fois la forme et la texture de visages à l'aide d'un nombre réduit de paramètres. Après quelques rappels sur le calcul d'un modèle actif d'apparence, nous détaillerons la construction d'un modèle hiérarchique, calculé sur une représentation multi-résolution de visages, à base de filtres de Gabor. Ce modèle peut être vu comme un intermédiaire entre le modèle actif de forme *(ASM)* et le modèle actif d'apparence. Dans le modèle d'apparence, la texture du visage est représentée par les valeurs de l'ensemble des pixels inclus dans l'enveloppe convexe de la forme du visage. Dans le modèle hiérarchique, la texture du visage n'est représentée qu'au niveau de quelques points intérieurs au visage à l'aide de bancs de filtres de Gabor. La représentation tient donc compte du voisinage de chacun des points sélectionnés, au travers de différents niveaux de résolution et selon différentes orientations. Les bancs de filtres, lorsqu'ils sont utilisés conjointement avec un maillage triangulaire de visages, ont jusqu'alors montré leur efficacité et leur robustesse pour la détection de traits et l'identification de personnes [9]. Dans le modèle actif de forme, la texture du visage n'est prise en compte que sous la forme de profils de niveaux de gris sur quelques segments de droites orthogonaux aux contours supposés de l'objet analysé. Les modèles *ASM, AAM* et *AAM* hiérarchique évoluent tous dans un champ de potentiel issu d'un apprentissage. Dans cette étude, nous évaluons le modèle hiérarchique par rapport à l'*AAM* proposé par Cootes et al., pour un problème de détection de la pose 2D et des traits caractéristiques de visages (yeux, bouche, etc.).

2 Le modèle actif d'apparence

Le modèle d'apparence est construit à partir d'exemples de visages constituant une base d'apprentissage, chaque visage étant décrit par une forme et une texture (luminance des pixels inclus dans la forme). Les formes de l'ensemble d'apprentissage sont recalées les unes par rapport aux autres par une transformation procrustéenne généralisée, et les textures sont normalisées puis alignées sur la forme moyenne des visages de l'ensemble d'apprentissage [2].
Considérons N formes, chacune composée de n points dans un espace de dimension 2 (le support de l'image). Chaque forme recalée est représentée par un vecteur $\mathbf{s} = [x_1, \ldots, x_n, y_1, \ldots, y_n]^T$ et chaque texture par un vecteur des niveaux de gris de M pixels, normalisés et alignés, $\mathbf{g} = [g_1, \ldots, g_M]^T$. Une analyse en composantes principales de chacune des deux classes de vecteurs permet de générer des formes et textures synthétiques à l'aide des équations suivantes : $\mathbf{s} = \bar{\mathbf{s}} + \Phi_s \mathbf{b}_s$ et $\mathbf{g} = \bar{\mathbf{g}} + \Phi_g \mathbf{b}_g$, où $\bar{\mathbf{s}}$, $\bar{\mathbf{g}}$ sont respectivement les vecteurs de forme et de texture moyens, Φ_s, Φ_g les matrices de vecteurs propres représentant les modes de variation, et $\mathbf{b}_s$, $\mathbf{b}_g$ les vecteurs de com-

posantes principales de forme et de texture. Une ACP supplémentaire est calculée sur les vecteurs $\mathbf{b}$ résultant de la concaténation des vecteurs $\mathbf{b}_s$ et $\mathbf{b}_g$ associés à chaque visage d'apprentissage [1] : $\mathbf{b}_s = \Phi_{c,s}\,\mathbf{c}$ et $\mathbf{b}_g = \Phi_{c,g}\,\mathbf{c}$, où $\mathbf{c}$ est le vecteur d'apparence de la classe de visages appris (il contrôle à la fois la forme et la texture d'un visage). Un nouveau visage peut ainsi être synthétisé à l'aide des équations $\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q}_s\,\mathbf{c}$ et $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q}_g\,\mathbf{c}$, où les matrices $\mathbf{Q}_s$ et $\mathbf{Q}_g$ décrivent les modes de variations présents dans l'ensemble d'apprentissage. L'apparence d'un visage et sa pose (position, orientation et échelle) dans l'image sont enfin représentés par le vecteur $\mathbf{p}^T = (\mathbf{c}^T \,|\, \mathbf{t}^T)$ où $\mathbf{t}$ contient les paramètres d'une transformation globale rigide.

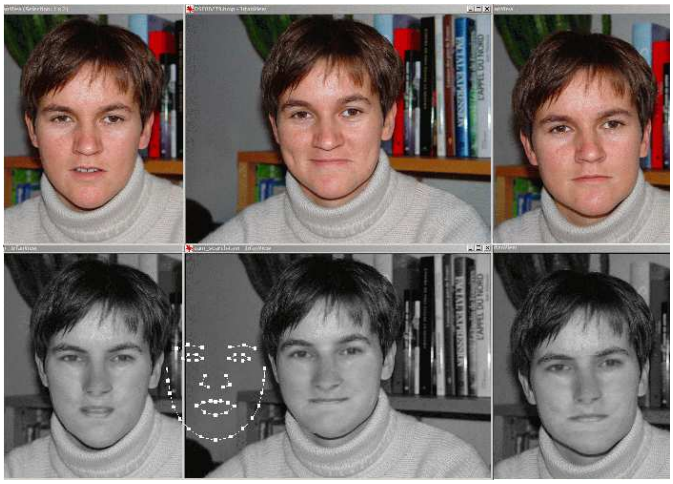


FIG. 1 – *En haut : visages originaux, inconnus du modèle. En bas : adaptation automatique du modèle actif d'apparence, calculé sur une base de 375 visages neutres et expressifs, en l'initialisant sur une position proche du visage original. Le visage synthétique est superposé au visage original. La forme moyenne $\bar{\mathbf{s}}$ du modèle est illustrée par les points blancs.*

FIG. 2 – *Adaptation itérative du modèle, en partant d'une pose et d'une apparence proche du visage cible situé en bas à droite.*

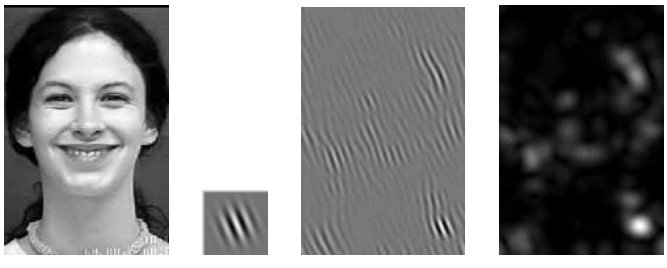
Cootes et al. proposent dans [2] une procédure d'adaptation automatique d'un modèle d'apparence à un visage cible inconnu, à l'aide d'un algorithme de descente de gradient, en supposant un gradient fixe à chaque itération, estimé au préalable par apprentissage : la relation $\delta\mathbf{p} = -\mathbf{R}\;\mathbf{r}(\mathbf{p})$ entre l'erreur de reconstruction $\mathbf{r}(\mathbf{p})$ (différence entre un vecteur de texture donné, extrait de l'image, et un vecteur généré par le modèle d'apparence) et la variation $\delta\mathbf{p}$ des paramètres du modèle est donc apprise au préalable, avec $\mathbf{R} = \left(\frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T \frac{\partial \mathbf{r}}{\partial \mathbf{p}}\right)^{-1} \frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T$.
Nous montrons dans la figure 1 les résultats de trois adaptations d'un modèle d'apparence, construit à partir de 375 visages d'apprentissage adultes, neutres et expressifs. Le vecteur d'apparence $\mathbf{c}$ contient dans ce cas 40 composantes, et le gradient définissant la matrice $\mathbf{R}$ est calculé par différenciation numérique [3, 8]. La figure 2 montre l'adaptation du même modèle à un visage jeune, inconnu.

3 Modèle d'apparence hiérarchique basé sur des filtres de Gabor

Dans le modèle d'apparence décrit dans la section 2, la recherche d'un visage s'appuie sur des valeurs de niveau de gris. Or, cette représentation rend l'algorithme sensible à une variation locale de luminosité. Les performances de détection peuvent se dégrader lorsque l'éclairage est trop différent de celui utilisé pendant l'apprentissage. Nous étudions ici l'apport d'une représentation par filtres de Gabor 2D dans un tel modèle, en notant que ces derniers, associés à une structure de graphe élastique, ont montré leur intérêt pour la détection et la reconnaissance de visages [9]. L'image $\mathcal{I}$ est décrite au voisinage d'un point $\mathbf{x} = [x,y]$, par ses réponses $\mathcal{J}_j(\mathbf{x}) = \mathcal{I} * \psi_j(\mathbf{x})$ à une famille (ψ_j) de filtres de Gabor. Chaque noyau $\psi_j(\mathbf{x}')$ code une onde plane caractérisée par le vecteur d'onde $\vec{k}_j$, et atténuée par une enveloppe Gaussienne (figure 3.b):

$$\psi_j(\mathbf{x}') = \frac{\|\vec{k}_j\|^2}{\sigma^2} \exp\left(-\frac{\|\vec{k}_j\|^2 \|\mathbf{x}'\|^2}{2\sigma^2}\right) \left[\exp(i\vec{k}_j \cdot \mathbf{x}') - \exp\left(-\frac{\sigma^2}{2}\right)\right]$$

Le vecteur d'onde $\vec{k} = 2\pi f(\cos\phi; i\sin\phi)^t$ définit la fréquence f et l'orientation ϕ de l'onde plane. Nous considérerons dans nos expériences la famille de $4 \times 8 = 32$ filtres, comportant 4 échelles ou fréquences $f_\nu = 2^{-(2+\frac{\nu}{2})}$, $\nu \in \{0,\ldots,3\}$, et 8 orientations $\phi_\mu = \mu\frac{\pi}{8}$, $\mu \in \{0,\ldots,7\}$, donc l'indice $j = \mu + 8\nu \in \{0,\ldots,31\}$. Cette discrétisation permet de couvrir de façon satisfaisante la bande de fréquences qui nous intéresse.



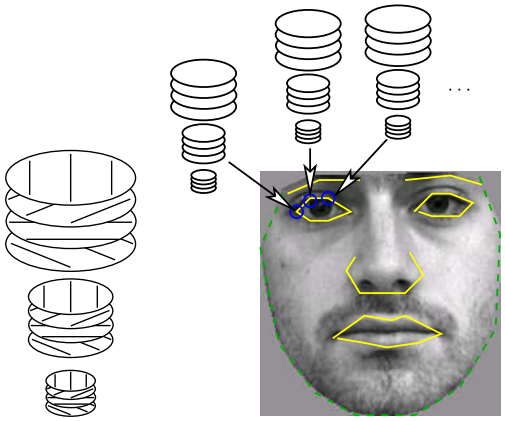
(a) (b) (c) (d)

FIG. 3 – *(a) Image originale. (b) Filtre $\psi_{\mu+8\nu}$ d'échelle $\nu = 1$, orienté à $\pi/8$ ($\mu = 1$), partie réelle. (c) Réponse du filtre, partie réelle. (d) Réponse du filtre, module.*

Comme le montrent les figures 3.a-d, lorsqu'on applique un filtre $\psi_{\mu+8\nu}$ d'échelle et d'orientation données, la réponse met en évidence des motifs précis dans l'image. De plus, le noyau $\psi_j(\mathbf{x}')$ étant de moyenne nulle, la réponse de l'image en un point $\mathbf{x}$ donné ne dépendra pas du niveau de gris moyen lo-

1. Remarque : le modèle pourrait tout aussi bien être calculé par une seule ACP de vecteurs issus de la concaténation de la forme et la texture des visages d'un ensemble d'appentissage.

cal en ce point. On peut donc espérer que cette représentation sera moins sensible que les niveaux de gris à une variation localisée d'éclairage. Nous utiliserons le module de la réponse $|\mathcal{J}_j(\mathbf{x})|$ (figure 3.d), qui évolue plus lentement que les autres composantes (phase ou partie réelle/imaginaire, cf. figure 3.c) : le module permet a priori de mieux détecter les structures d'intérêt, notamment lorsque la position initiale est peu éloignée de la position optimale.



(a) (b)

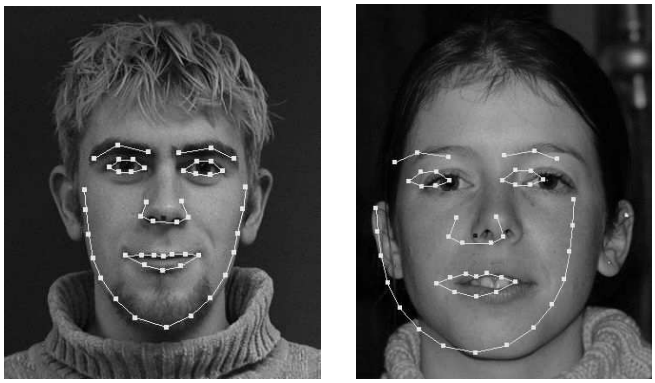
FIG. 4 – *(a) Représentation schématique d'un banc de filtres, ici de 3 échelles $\times$ 4 orientations (dans les expériences, on emploie des bancs de $4 \times 8 = 32$ filtres). (b) Les réponses aux filtres seront calculées en chaque point annoté intérieur au visage, sur la texture de niveaux de gris ramenée à la forme moyenne des visages de l'ensemble d'apprentissage.*

Ainsi, l'idée proposée consiste à calculer un modèle d'apparence comme présenté dans la section 2, mais en remplaçant le vecteur de texture $\mathbf{g}$ par les réponses de l'image aux 32 filtres de Gabor. La mise en œuvre de cette idée repose sur trois points, illustrés par la figure 4 :

1. Les 32 réponses ne sont pas calculées en chaque pixel, mais uniquement aux points annotés, composant le vecteur $\mathbf{s}$: en effet, ces points annotés correspondent aux traits marqués et stables du visage, on peut donc les considérer comme les lieux portant le plus d'information relative à un visage.
2. Afin de compenser les variations de pose du visage (échelle, rotation dans le plan), il convient de calculer les réponses sur une texture qui soit ramenée à la forme de référence.
3. Il nous semble préférable de ne calculer les réponses que sur les points intérieurs du visage. En effet, si on utilise les réponses au bord du visage, on prendra en compte un fond qui est par nature imprévisible donc difficile à modéliser. Ou bien on doit remplacer le fond par des valeurs arbitraires de niveaux de gris ; mais on risque alors d'induire un artefact dans le modèle, par exemple en créant un contour très marqué au bord, qui domine les autres réponses.

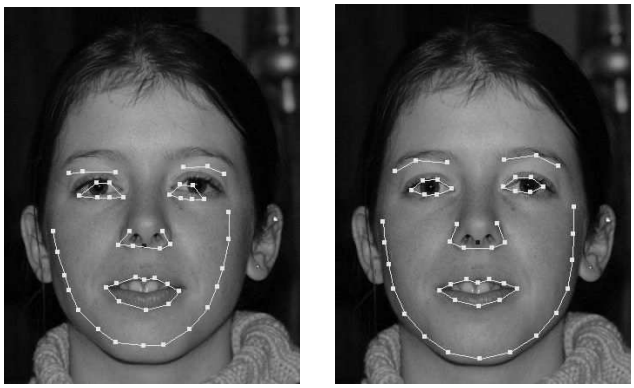
4 Résultats

La figure 5.a montre un exemple d'image d'apprentissage issu d'une base de 37 visages neutres [8]. Pour le problème de localisation des points du visage, nos expériences visent à comparer les performances de 2 méthodes : le modèle actif d'apparence de niveaux de gris décrit dans la section 2, et celui utilisant les réponses de filtres de Gabor. D'après les résultats obtenus, le modèle basé sur les filtres de Gabor se montre plus robuste que celui basé sur les niveaux de gris, lorsque les conditions d'illumination de test sont très différentes de conditions apprises. Ceci est illustré par les détections des figures 6, obtenues sur une image issue d'une autre base, en partant de la position initiale de la figure 5.b. Lorsque les conditions d'illuminations sont semblables, le modèle hiérarchique se montre moins précis que le modèle classique. En considérant différents changements d'échelle, translations, ou rotations, l'erreur de détection moyenne est environ de 2 pixels, contre 1,6 pixels pour le modèle classique.



(a) (b)

FIG. 5 – *(a) Exemple d'image d'apprentissage annotée. (b) Sur une image inconnue extraite d'une autre base, initialisation de la détection avec la forme moyenne $\bar{\mathbf{s}}$, positionnée avec une erreur de 25 pixels en x.*



(a) (b)

FIG. 6 – *(a) Détection par l'AAM classique. (b) Détection par l'AAM hiérarchique (AAM + Gabor).*

Le modèle hiérarchique se révèle efficace également lorsqu'il est initialisé loin du visage cible. La figure 7 montre l'adaptation des modèles classique et hiérarchique à un visage cible inconnu, lorsque l'initialisation est éloignée de 42 pixels de la vraie position. Étant donné que seuls les points annotés internes au visage sont pris en compte pour le calcul du modèle hiérar-

chique, la robustesse de celui-ci pourrait être réduite ; les résultats montrent que la localité du modèle hiérarchique (en terme de points annotés sur les yeux, nez et bouche sans le pourtour du visage) est compensée par le fait qu'à chacun des points sont associées des réponses de Gabor selon différentes orientations et surtout différents niveaux de résolution.

FIG. 7 – *De gauche à droite : Sur une nouvelle image inconnue, initialisation de la détection avec la forme moyenne* $\bar{\mathbf{s}}$*, positionnée avec une erreur de 30 pixels vers la droite et 30 pixels vers le bas ; Détection obtenue avec l'AAM classique ; Détection obtenue avec l'AAM hiérarchique (Gabor).*

Enfin, des résultats expérimentaux montrent que la précision du modèle d'apparence hiérarchique pour la détection des traits de visages reste correcte lorsque la taille de la base d'apprentissage servant au calcul du modèle diminue ; ceci n'est pas vérifié dans le cas du modèle classique. Cette observation est faite en ne prenant que 10 ou 20 images d'apprentissage pour construire les modèles.

5 Conclusion

En partant du cadre formel des modèles actifs d'apparence proposés par Cootes [2], nous proposons dans cet article une méthode nouvelle de construction d'un modèle d'apparence hiérarchique, basé sur des filtres de Gabor. Ce travail se place dans la problématique de la détection des traits de visages dans une image fixe. Le modèle d'apparence hiérarchique présente l'avantage de rendre la détection plus robuste aux variations locales de luminosité.
Si on cherche à générer des visages réalistes, le modèle hiérarchique ne fournit pas un cadre adapté : comme les noyaux de Gabor sont de moyenne nulle, les réponses de Gabor ne permettent pas de reconstituer le niveau de gris moyen local [7]. Pour cette tâche de synthèse de visages, des travaux en cours montrent l'intérêt du modèle classique pour modifier l'expression faciale d'une personne, après filtrage de son expression initiale [1]. Le modèle hiérarchique peut quant à lui être utile pour la reconnaissance des gestes faciaux (expressions faciales, mimiques, clignements des yeux, etc.). En effet, la reconnaissance peut se faire aussi bien à partir des vecteurs d'apparence calculés selon la méthode classique [1], qu'en utilisant ceux fournis par la méthode hiérarchique.

Remerciements

Ce travail s'est effectué dans le cadre d'une ACI Jeunes Chercheurs du Ministère de la Recherche. Les auteurs remercient également Mademoiselle Coralie BONNET, étudiante à l'UTC, pour sa participation au projet et son aide à la programmation et aux tests des modèles.

Références

[1] B. Abboud, F. Davoine, and M. Dang. Expressive face recognition and synthesis. In *IEEE Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction (in conjunction with CVPR)*, Madison, U.S.A., June 2003.

[2] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[3] T.F. Cootes and P. Kittipanya-ngam. Comparing variations on the active appearance model algorithm. In *British Machine Vision Conference*, pages 837–846, Cardiff University, September 2002.

[4] T. Ezzat and T. Poggio. Facial analysis and synthesis using image-based models. In *International Conference on Automatic Face and Gesture Recognition*, pages 116–121, 1996.

[5] S.B. Gokturk, J.-Y. Bouguet, and R. Grzeszczuk. A data-driven model for monocular face tracking. In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[6] M. Malciu. *Approches orientées modèle pour la capture des mouvements du visage en vision par ordinateur*. Thèse de doctorat, Université René Descartes, Paris V, INT, Unité de Projets Artemis, décembre 2001.

[7] Michael Pötzsch, Thomas Maurer, Laurenz Wiskott, and Christoph von der Malsburg. Reconstruction from graphs labeled with responses of gabor filters. In *International Conference on Artificial Neural Networks*, pages 845–850, Bochum, Germany, 1996.

[8] M.B. Stegmann. Active appearance models: Theory, extensions and cases. In *Master Thesis, IMM-EKS-2000-25*, Technical University of Denmark, Lyngby, 2000.

[9] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. In L. C. Jain, U. Halici, I. Hayashi, and S. B. Lee, editors, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, chapter 11, pages 355–396. CRC Press, 1999.

[10] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.