# Analyse de visages et d'expressions faciales par modèle actif d'apparence

## Face and facial expression analysis based on an active appearance model

**Franck Davoine, Bouchra Abboud et Van Mô Dang**

HEUDIASYC, UMR 6599 CNRS, Université de Technologie de Compiègne, BP 20529, 60205 Compiègne cedex, France.
prenom.nom@hds.utc.fr

Manuscrit reçu le 9 février 2004

Résumé et mots clés

Dans cet article, nous nous intéressons à l'extraction automatique des traits de visages (yeux, sourcils, nez, bouche, menton) ainsi qu'à la reconnaissance des six expressions faciales définies par Ekman [19]. Nous exploitons pour cela des versions modifiées du modèle actif d'apparence initialement proposé par Cootes *et al.* [11] qui permet de représenter à la fois la forme et la texture d'un visage. L'extraction des traits faciaux est faite à l'aide d'un modèle actif d'apparence hiérarchique, calculé à partir des réponses de visages à des bancs de filtres de Gabor. Deux modèles d'expressions faciales sont ensuite proposés, calculés à patir du modèle d'apparence standard (non hiérarchique), pour reconnaître puis supprimer ou modifier l'expression d'un visage inconnu.

Visages, traits caractéristiques, reconnaissance, filtrage et modification d'expressions faciales, modèle actif d'apparence, ACP, filtres de Gabor, regression.

Abstract and key words

In this paper, methods are proposed for facial feature detection (eyes, brows, nose, mouth, chin) and for facial expression recognition. The methods are based on modified versions of the standard Active Appearance Model proposed by Cootes *et al.* [11] to control both the shape and the texture of a given face. The detection algorithm makes use of an active appearance model computed on hierarchical Gabor descriptions a set of training faces. In a second part, two expression models are proposed, based on the standard AAM, and used to recognize and then to cancel or modify the facial expression of a given unknown face.

Faces, facial features, facial expression cancellation, modification or recognition, active appearance model, PCA, Gabor filters, regression.

# 1. Introduction

L'analyse de visages par traitement d'images est encore aujourd'hui un sujet de recherche très actif puisqu'il concerne de nombreux domaines d'application tels que par exemple la sécurité (biométrie, surveillance), la robotique (interaction homme-machine, *affective computing*), le handicap (communication par le visage), les jeux vidéo ou les télécommunications à très bas débits (clones synthétiques). Les recherches englobent la détection, le suivi, le codage, la reconnaissance et la synthèse de visages en tenant compte des variations possibles de leur apparence (pose tridimensionnelle, regard, lèvres, expressions, âge, genre, mouvements faciaux et comportement facial, occultations, etc.). Parmi les méthodes proposées, nombreuses sont celles qui utilisent des modèles permettant une coopération entre l'analyse et la synthèse d'un visage [54, 23, 36, 24, 37]. Dans cet article, nous nous intéressons plus particulièrement au modèle actif statistique d'apparence *(AAM)*, initialement proposé par Cootes *et al.* [11], et qui permet de contrôler à la fois la forme et la texture de visages à l'aide d'un nombre réduit de paramètres. Nous présentons deux utilisations possibles du modèle pour (i) extraire automatiquement les traits caractéristiques d'un visage vu de face et (ii) reconnaître et synthétiser des expressions faciales.

Des travaux de recherche en psychologie ont démontré que les expressions faciales jouent un rôle prépondérant dans la coordination de la conversation humaine [6], et ont un impact plus important sur l'auditeur que le contenu textuel du message exprimé. Mehrabian [38] remarque que la contribution du contenu textuel d'un message verbal en « face à face » à son impact global se limite à 7 % alors que les signaux conversationnels (accentuation de mots, ponctuation, marqueurs d'une question, indicateurs d'une recherche de mots, etc.) et l'expression faciale du locuteur contribuent respectivement à 38 % et 55 % de l'impact global du message exprimé. Par conséquent, l'expression faciale peut être considérée comme une modalité essentielle de la communication humaine.

L'analyse automatique des expressions faciales constitue un outil important pour la recherche dans les domaines de l'étude du comportement et de la psychologie, ainsi que dans les domaines de la compression d'images et de l'animation de visages synthétiques [43]. Elle repose fréquemment sur le système *FACS*, proposé par Ekman et Friesen en 1978 [20], et qui constitue une description objective de signaux faciaux décrits par 46 mouvements élémentaires indépendants ou « unités actions faciales ». Ekman a également montré que les expressions faciales de six catégories émotionnelles de base sont universellement reconnues, à savoir : la colère, le dégoût, la peur, la joie, la surprise et la tristesse [19].

Dans le passé, les travaux de recherche sur l'analyse des expressions faciales se situaient principalement dans le cadre de la psychologie [8]. Les progrès effectués dans des domaines connexes tels que la détection, le suivi et la reconnaissance de visages [55] ont apporté une contribution significative à la recherche dans le domaine de l'analyse, de la synthèse et de la reconnaissance d'expressions faciales [14, 44, 49, 21]. Nous listons dans ce chapitre d'introduction, sans volonté d'être exhaustifs, une sélection de travaux relatifs à l'analyse et la synthèse de visages.

#### Détection et analyse des traits faciaux

Avant de procéder à l'analyse de l'expression faciale d'un visage fixe ou en mouvement, il convient de le détecter ou de le suivre afin d'en extraire des informations pertinentes. Plusieurs méthodes de détection sont décrites dans [54, 27]. Selon le cas, elles exploitent une représentation globale ou locale du visage, codée par exemple sous la forme de vecteurs de couleurs ou de niveaux de gris de pixels, de vecteurs de mouvement ou de réponses à différents filtres (ondelettes, Gabor, etc.). Les vecteurs de visages étant de grande taille, ils sont souvent transformés à l'aide de méthodes linéaires de réduction de dimension telles que l'analyse en composantes principales (ACP) ou indépendantes (ACI). Lorsqu'en plus, les vecteurs exhibent des caractéristiques non linéaires (dues par exemple à des variations d'éclairage ou d'orientation dans l'espace), ils peuvent être transformés à l'aide de méthodes non-linéaires telles que l'analyse en composantes principales à noyau [47, 34]. Les méthodes locales permettent une modélisation du visage dans les régions susceptibles de se modifier selon par exemple les expressions faciales affichées. Viola *et al.* [51] ont récemment proposé une méthode de détection de visages très rapide et compétitive en terme de taux d'erreurs par rapport aux méthodes concurrentes. Les auteurs exploitent un codage multirésolution de l'image, connu sous le nom d'*image intégrale* et obtenu par filtrage. Une variante de l'algorithme *AdaBoost* leur permet de sélectionner un faible de nombre de caractéristiques faciales, à partir d'exemples de visages et de contre-exemples, de façon à entraîner un jeu de classifieurs. Les visages présents dans une image sont ensuite détectés à l'aide d'une cascade de ces classifieurs.

Dans le but de représenter le mouvement intérieur au visage, lorsque celui-ci est détecté dans l'image, Black *et al.* [4] utilisent des modèles locaux paramétriques. Ils estiment le mouvement relatif des traits faciaux dans le repère du visage. Les paramètres de ce mouvement servent par la suite à représenter l'expression faciale. De manière similaire, Cohn *et al.* [10] utilisent un algorithme hiérarchique pour effectuer le suivi des traits caractéristiques par estimation du flot optique. Les vecteurs de déplacement représentent l'information sur les changements d'expression faciale. Padgett *et al.* [42] utilisent des gabarits d'œil et de bouche, calculés par analyse en composantes principales d'un ensemble d'apprentissage, en association avec des réseaux de neurones. D'autre part, Hong *et al.* [28] utilisent un modèle global basé sur des graphes étiquetés construits à partir de points de repère distribués sur le visage. Les noeuds de ces graphes sont formés par des vecteurs dont chaque élément est la

réponse à un filtrage de Gabor extraite en un point donné de l'image. Finalement, Cootes *et al.* [11] utilisent une représentation par modèle actif d'apparence *(AAM)* pour extraire automatiquement des paramètres caractérisant un visage.

#### Reconnaissance d'expressions faciales

Un grand nombre de systèmes d'analyse d'expressions faciales proposés dans la littérature visent à reconnaître et à mesurer l'amplitude d'unités d'actions faciales à partir de visages vus de face, fixes ou en mouvement. D'autres système cherchent plutôt à reconnaître un ensemble limité d'expressions « prototypes » telles que la joie, la colère, le dégoût, la tristesse, la peur, la surprise, ou d'autres actions telles qu'un clignement d'œil ou un cri. Dans le cadre de cet article, nous nous intéresserons plus particulièrement à cette deuxième catégorie de méthodes.

La reconnaissance d'un nombre pré-défini d'expressions faciales nécessite au préalable le choix d'une représentation parcimonieuse des visages permettant l'émergence de classes distinctes d'expressions dans un ensemble d'apprentissage. Différents méthodes ont été proposées, exploitant par exemple l'analyse en composantes principales [41] ou indépendantes ou l'analyse discriminante, linéaire ou non-linéaire [25]. Vient ensuite le choix d'une méthode de classification s'appuyant sur des mesures de distances déterministes ou probabilistes entre individus [39], sur des machines à vecteur de support [1] ou des réseaux de neuronaux.

Afin de reconnaître une expression faciale comme l'une des six expressions universelles définies par Ekman [19] auxquelles s'ajoute l'expression neutre, Hong *et al.* [28] partent du principe que deux personnes qui se ressemblent affichent la même expression de manière similaire. Un graphe étiqueté est attribué à l'image de test puis la personne connue la plus proche est déterminée à l'aide d'une méthode de mise en correspondance de graphes élastiques. La galerie personnalisée de cette personne est alors utilisée pour reconnaître l'expression faciale de l'image de test. Un graphe étiqueté par des réponses de filtres de Gabor est par ailleurs utilisé par Lyons *et al.* [35]et Bartlett *et al.* [2]. L'ensemble des graphes construits sur un ensemble d'apprentissage est ensuite soumis à une ACP puis analysé à l'aide d'une analyse discriminante linéaire (ADL) afin de séparer les vecteurs dans des classes ayant des attributs faciaux différents. Le graphe étiqueté de l'image testée sera alors projeté sur les vecteurs discriminants de chaque classe afin de déterminer son éventuelle appartenance à cette classe. Dans une finalité identique, Essa et Pentland [21] extraient des gabarits spatio-temporels de l'énergie du mouvement du visage pour chaque expression faciale. Le critère de similarité repose sur la distance euclidienne entre ces gabarits et l'énergie du mouvement de l'image observée. Heisele *et al.* [26] utilisent des machines à vecteur de support *(SVM)* dans le cadre de la reconnaissance de visages par des méthodes globales ainsi que par des méthodes reposant sur des traits caractéristiques. De manière identique, l'algorithme de reconnaissance de visages *FaceIt* est basé sur une analyse locale des traits caractéristiques *(LFA)* développée par Penev et Atick [45]. Draper *et al.* [15] comparent les performances de l'analyse en composantes principales et de l'analyse en composantes indépendantes pour la reconnaissance de visages et d'expressions faciales en se basant sur le codage *FACS* [20]. Yang propose dans [53] une analyse en composantes principales à noyau pour la reconnaissance de visages. Finalement, Edwards *et al.* [18] utilisent le modèle actif d'apparence pour reconnaître l'identité d'un individu observé de manière robuste par rapport à l'expression faciale ainsi que l'illumination et la pose. Pour ceci, le critère de similarité utilisé repose sur la distance de Mahalanobis, et une ADL est appliquée afin de maximiser la séparation des classes.

#### Synthèse d'expressions faciales

La synthèse d'expressions faciales est une tâche difficile compte tenu de la complexité de la forme et de la texture des visages. De plus, le visage présente des rides et des plis ainsi que d'autres variations subtiles de forme et de texture qui ont une importance cruciale dans la compréhension et la représentation des expressions faciales. Dans cette perspective, les techniques d'interpolation et de déformation offrent une approche intuitive pour l'animation de visages. Plusieurs travaux visent à traiter séparément la texture et la forme d'un visage [3, 5, 50]. Pighin *et al.* [46] utilisent des techniques de *morphing* 2D combinées avec des transformations d'un modèle géométrique 3D, pour créer des modèles faciaux réalistes tridimensionnels à partir de photographies, et pour construire des transitions lisses entre les différentes expressions faciales. Dans la même optique, Blanz *et al.* [5] déforment d'un modèle géométrique 3D, sur lequel est projetée la numérisation 3D de la texture d'un visage. En outre, dans le cadre du logiciel *Video-Rewrite*, Bregler *et al.* [7] associent des techniques de suivi de points 2D de la bouche d'un orateur dans une séquence d'apprentissage à des techniques de *morphing* pour animer les lèvres d'une personne inconnue prononçant les même paroles. Dans une finalité analogue, Ezzat *et al.* [22] utilisent une représentation par modèle déformable multidimensionnel et une technique de synthèse de trajectoire pour contrôler les mouvements de la bouche d'un visage parlant. Cette représentation permet de synthétiser des configurations inconnues de lèvres parlantes « vidéo-réalistes » à partir d'une séquence vidéo initiale. Chuang *et al.* [9] utilisent quant à eux une ACP combinée à un modèle bilinéaire pour synthétiser une nouvelle expression sur un visage parlant. Finalement, Kang *et al.* [32] utilisent le modèle actif d'apparence combiné avec une régression linéaire pour annuler l'expression faciale d'un visage dans le but d'améliorer les performances d'un algorithme de reconnaissance de visages.

**Dans cet article**, nous décrivons la construction d'un modèle actif d'apparence hiérarchique calculé sur une représentation multirésolution de visages, à base de filtres de Gabor. Ce modèle peut être vu comme un intermédiaire entre le modèle actif de

forme *(ASM)* [33] et le modèle actif d'apparence [11]. Dans le cas du modèle d'apparence, la texture du visage est représentée par les valeurs de l'ensemble des pixels inclus dans l'enveloppe convexe des points de la forme du visage. Dans le cas du modèle hiérarchique proposé ici, la texture n'est représentée qu'au niveau de quelques points intérieurs au visage à l'aide de bancs de filtres de Gabor. La représentation tient ainsi compte du voisinage de chacun des points sélectionnés, au travers de différents niveaux de résolution et selon différentes orientations, contrairement au cas de l'*ASM* pour lequel 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 bancs de filtres de Gabor, lorsqu'ils sont utilisés conjointement avec un maillage triangulaire de visages, ont en outre montré leur efficacité et leur robustesse pour la détection de traits faciaux et l'identification de visages [52]. Dans cet article, nous évaluons l'intérêt du modèle hiérarchique par rapport à l'*AAM*} proposé par Cootes *et al.* [11], pour un problème de détection de la pose 2D et des traits caractéristiques de visages (yeux, bouche, etc.). Dans une deuxième partie, nous proposons une approche originale de la reconnaissance d'expressions faciales basée sur le modèle d'apparence standard de Cootes *et al.* (non hiérarchique). Nous présentons une extension de la méthode décrite dans [32] pour annuler l'expression d'un visage, à une application de synthèse de nouvelles expressions faciales. Enfin nous introduisons une nouvelle méthode d'interpolation pour la synthèse et l'annulation d'expressions faciales.

# 2. Modèle actif d'apparence

Cette section donne une description rapide du modèle actif d'apparence. Sa construction est d'abord expliquée et quelques résultats expérimentaux montrant l'adaptation du modèle sur des visages inconnus sont ensuite donnés. Nous vérifions également l'efficacité d'une variante du modèle d'apparence, calculée à partir d'une seule ACP.

### 2.1 Description

Il a déjà été démontré [11] que le modèle actif d'apparence est un outil puissant permettant de représenter des visages de façon réaliste et naturelle. Il se base sur une technique d'ACP permettant de représenter conjointement les variations de forme et de texture présentes dans un ensemble d'apprentissage. En effet, après avoir aligné toutes les formes de l'ensemble d'apprentissage par rapport à la forme moyenne à l'aide d'une analyse Procrustéenne [16], le modèle statistique de forme est donné par :

$$\mathbf{s}_i = \bar{\mathbf{s}} + \Phi_s \mathbf{b}_{si}, \tag{1}$$

où $\mathbf{s}_i$ est la forme reconstruite, $\Phi_s$ est une matrice contenant les principaux modes de variation de forme, et $\mathbf{b}_{si}$ est un vecteur contrôlant la forme reconstruite. Il est alors possible de déformer les textures de l'ensemble d'apprentissage sur la forme moyenne $\bar{\mathbf{s}}$ pour séparer la texture de la forme. De manière similaire, après avoir calculé la texture sans forme moyenne $\bar{\mathbf{g}}$ et normalisé toutes les textures de l'ensemble d'apprentissage par rapport à la texture moyenne à l'aide d'une translation et d'une mise à l'échelle des niveaux de gris, le modèle statistique de texture est donné par :

$$\mathbf{g}_i = \bar{\mathbf{g}} + \Phi_t \mathbf{b}_{ti}, \tag{2}$$

\noindent goù $\mathbf{g}_i$ est la texture sans forme reconstruite, $\Phi_t$ est une matrice contenant les modes principaux de variation de texture dans l'ensemble d'apprentissage, et $\mathbf{b}_{ti}$ est un vecteur contrôlant la texture reconstruite, sans forme.
Après pondération des valeurs des composantes de $\mathbf{b}_{si}$ avec des poids convenablement choisis pour rendre leur ordre de grandeur comparable à $\mathbf{b}_{ti}$ [11], ces vecteurs sont concaténés. En pratique les composantes de $\mathbf{b}_{si}$ sont pondérées par un poids égal à la racine carrée du rapport de la somme des valeurs propres associées à $\Phi_t$ sur la somme des valeurs propres associées à $\Phi_s$. Une ACP supplémentaire sur les nouveaux vecteurs obtenus retourne le modèle statistique d'apparence donné par :

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \tag{3}$$

$$\mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i, \tag{4}$$

où $Q_s$ et $Q_t$ sont des matrices codant les principaux modes de variation d'apparence dans l'ensemble d'apprentissage, et $\mathbf{c}_i$ est un vecteur de paramètres d'apparence permettant de contrôler simultanément la forme et la texture du visage. Cette troisième ACP s'avère utile puisque les vecteurs $\mathbf{b}_{si}$ et $\mathbf{b}_{ti}$ sont corrélés.
Une autre manière de calculer le modèle d'apparence consiste à concaténer, pour chaque visage $i$ de l'ensemble d'apprentissage, les vecteurs de forme $\mathbf{s}_i$ et de texture $\mathbf{g}_i$ pour former le vecteur $\mathbf{b}_{sti}$. Les vecteurs $\mathbf{s}_i$ sont multipliés par une matrice de poids $W_s$ convenablement choisie afin de rendre leur ordre de grandeur comparable à $\mathbf{g}_i$. En pratique, les vecteurs $\mathbf{s}_i$ sont pondérées par un poids égal au rapport de l'écart-type des valeurs des pixels normalisées par l'écart-type des formes normalisées de l'ensemble d'apprentissage.
En appliquant une ACP aux vecteurs $\mathbf{b}_{sti}$, le modèle d'apparence modifié est donné par :

$$\begin{pmatrix} \mathbf{W}_s \\ \mathbf{s}_i \mathbf{g}_i \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{s}} \\ \bar{\mathbf{g}} \end{pmatrix} + \mathbf{Q}_{st} \mathbf{c}'_i, \tag{5}$$

où $\mathbf{Q}_{st}$ est une matrice représentant les principaux modes de variation des vecteursp combinés de forme et de texture, et $\mathbf{c}'_i$ est

*Figure 1. Effets des perturbations associées aux six principaux modes de déformation d'un visage. Sur chaque ligne indexée par $j \in [1,6]$, perturbation du paramètre $c_j$ de* $\mathbf{c}$ *de :* $-3$, $-1{,}5$, $0$, $1{,}5$, *et* $3$ *fois son écart-type. Colonne centrale : visage moyen défini par* $\overline{\mathbf{s}}$ *et* $\overline{\mathbf{g}}$ *(*$\mathbf{c} = 0$*).*

le vecteur d'apparence contrôlant conjointement la forme et la texture reconstruite par le modèle. Ce dernier modèle présente l'intérêt de nécessiter le calcul d'une seule ACP, au lieu de trois. Nous constatons que les représentations simplifiées de visages obtenues à l'aide de ces deux modèles (à trois ou une ACP) sont très comparables, lorsque le même pourcentage de variabilité totale de l'ensemble d'apprentissage est conservé. Nous comparerons dans la section 4 les performances des deux modèles pour le problème de la reconnaissance d'expressions faciales.

Disposant du vecteur $\mathbf{c}_i$ (ou $\mathbf{c}'_i$), la forme et la texture du visage $i$ peuvent être retrouvées à partir des équations (3) et (4) (ou (5) pour la représentation modifiée). La texture sans forme reconstruite est alors déformée sur la forme reconstruite afin d'obtenir l'apparence globale du visage reconstruit. De plus, il est nécessaire d'ajouter au vecteur d'apparence $\mathbf{c}_i$, les paramètres d'une transformation géométrique permettant de contrôler la pose du visage reconstruit dans l'image. Ces paramètres sont classiquement ceux d'une similitude plane (rotation, facteur d'homothétie et translation), définie par le vecteur de pose $\mathbf{t}_i$.

En partant du principe que le vecteur $\mathbf{p}^{\mathbf{T}} = (\mathbf{c}^{\mathbf{T}}|\mathbf{t}^{\mathbf{T}})$ constitue une représentation d'un visage, le modèle actif d'apparence *AAM* peut s'adapter de manière itérative à un visage cible inconnu [12] afin de minimiser une image résiduelle $\mathbf{r}(\mathbf{p})$ qui n'est autre que la différence entre la texture du visage reconstruit et la texture qu'il recouvre dans l'image cible. La procédure d'adaptation de type « descente de gradient », décrite dans [12], repose sur la relation linéaire $\delta\mathbf{p} = -\mathbf{R}\,\mathbf{r}(\mathbf{p})$ qui permet, pour un résidu courant $\mathbf{r}(\mathbf{p})$, de calculer l'incrément $\delta\mathbf{p}$ à appliquer au vecteur $\mathbf{p}$ afin de minimiser l'erreur $|\mathbf{r}(\mathbf{p}+\delta\mathbf{p})|^2$. La matrice $\mathbf{R}$ est donnée par :

$$\mathbf{R} = \left( \frac{\partial \mathbf{r}^{\mathbf{T}}}{\partial \mathbf{p}} \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \right)^{-1} \frac{\partial \mathbf{r}^{\mathbf{T}}}{\partial \mathbf{p}} \tag{6}$$

\noindent La matrice $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ est constante, estimée par différentiation numérique à partir d'un ensemble d'apprentissage : cet ensemble est constitué de vecteurs $\mathbf{p}_j$ connus (pose du modèle adaptée manuellement à chaque visage $j$ de l'ensemble d'apprentissage), et la matrice $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ est calculée à partir des résidus engendrés par la perturbation systématique des composantes des vecteurs $\mathbf{p}_j$.

### 2.1. Résultats expérimentaux

Le modèle est calculé à partir de 375 visages extraits principalement de la base CMU [31] et représentant six classes d'expressions (colère, dégoût, peur, joie, surprise et tristesse) d'intensités fortes et modérées, ainsi que l'expression neutre. 53 points caractéristiques ont été manuellement positionnés sur ces visages par une même personne, en s'appuyant sur l'information de texture (les contours des traits caractéristiques). 375 vecteurs de texture composés de 5 871 pixels en ont été extraits. L'expérience nous montre que la précision avec laquelle sont extraits les points caractéristiques des visages affecte peu la qualité de l'adaptation du modèle à un visage cible. Dans [29], les auteurs calculent un modèle actif d'apparence facial en prédisant linéairement la forme à partir de la texture du visage. Dans [12], les auteurs vérifient que la précision de l'adaptation de la forme de ce modèle à un visage cible reste très proche, voire meilleure que celle obtenue à l'aide du modèle standard. Ceci nous conduit à conclure que le positionnement précis des points caractéristiques n'est pas une condition majeure pour la convergence du modèle d'apparence, au moins pour ce qui concerne la modélisation des expressions faciales à des fins par exemple de synthèse ou de classification (comme nous le verrons dans la section 4). Ceci serait différent si l'objectif était d'utiliser le modèle d'apparence pour une segmentation précise d'images médicales.

Le modèle standard est construit en conservant 50 modes de forme, 170 modes de texture et finalement 120 modes d'apparence. Le modèle modifié est quant à lui construit en gardant 170 modes d'apparence conservant ainsi pour les deux représentations 98 % de pla variabilité totale des vecteurs combinés.

Les adaptations à un visage inconnu du modèle d'apparence dans sa version standard (équations 3 et 4) et modifiée (équation 5) sont illustrées sur les figures 2.c et 2.d. Les figures 3 et 4 illustrent l'adaptation d'un modèle d'apparence de plus haute résolution composé, dans ce cas, de 13 085 pixels.

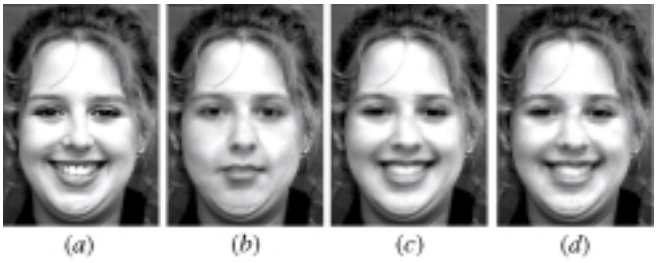


*Figure 2. a : Visage inconnu. b : initialisation du modèle (forme et texture moyenne). c : adaptation du modèle actif d'apparence. d : adaptation du modèle modifié n'impliquant qu'une ACP*

*Figure 3. Adaptation itérative du modèle actif d'apparence, en partant d'une pose et d'une apparence proches du visage cible situé en bas à droite.*

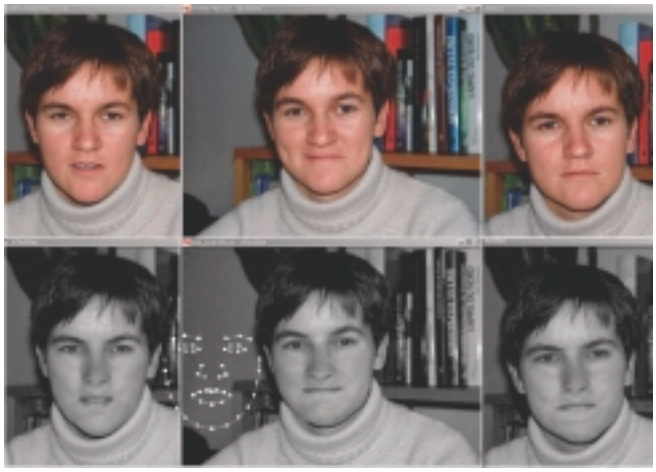

*Figure 4. En haut : trois visages originaux, inconnus. En bas : adaptations automatiques d'un modèle actif d'apparence, calculé sur une base de 375 visages neutres et expressifs. Dans les trois cas, la pose initiale du modèle est proche de celle du visage original. Le visage synthétique obtenu après convergence est superposé au visage original. (La forme moyenne $\overline{\mathbf{s}}$ du modèle est représentée par les points blancs)*

# 3. Modèle d'apparence hiérarchique

Dans le cas du modèle d'apparence décrit dans la section 2, la recherche d'un visage s'appuie sur des valeurs de niveaux 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. Dans [30], les auteurs proposent de remplacer, dans une variante de l'*AAM*, le modèle statistique de texture par un réseau d'ondelettes de Gabor. L'apparence d'un visage est ainsi partiellement représentée à l'aide d'un jeu d'ondelettes de Gabor, pré-sélectionnées par apprentissage et optimisation de type Levenberg-Marquardt. Le résidu de l'approximation d'un visage par un réseau d'ondelettes sert dans ce cas à prédire par régression linéaire les paramètres de forme et de pose du modèle d'apparence. La méthode permet d'adapter un modèle de forme à un visage, de façon robuste face à des occultations locales ou des variations d'éclairage. Les filtres de Gabor, associés à une structure de graphe élastique, ont en outre montré leur intérêt pour la détection robuste et la reconnaissance de visages [52].

Nous proposons dans cette section de calculer le modèle actif d'apparence sur des visages préalablement filtrés à l'aide d'ondelettes de Gabor, afin d'adapter la forme du modèle à un visage cible. Les deux méthodes d'adaptation, reposant sur une apparence « pixellique » (*cf.* section 2) ou multirésolution, pourront ainsi être comparées directement.

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 $(\psi_j)$ de filtres de Gabor ajustables en orientation et en bande de fréquence. 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 5.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 $\phi$ 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.

Comme le montrent les 5.a à 5.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 $\mathbf{x}$

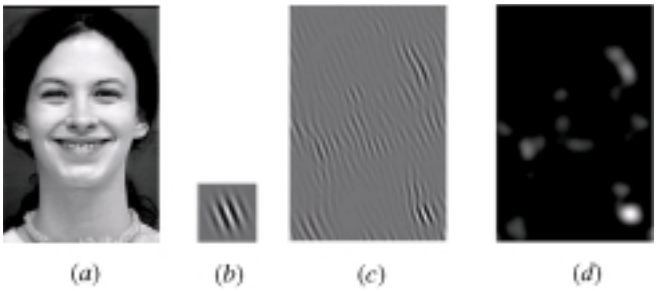


*Figure 5. (a) Image originale. (b) Filtre $\psi_{\mu+8\nu}$ d'échelle $\nu = 1$, orienté à $\pi/8$ ($\mu = 1$), partie réelle. (c) Réponse du filtre, partie réelle. (d) Réponse du filtre, module*

donné ne dépendra pas du niveau de gris moyen local 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 n'utiliserons que le module de la réponse $|\mathcal{J}_j(\mathbf{x})|$ (figure 5.d), sans tenir compte de l'information de phase : le module de l'image filtrée fournit une information locale sur la taille des détails de l'image (décomposition fréquentielle) et sur leur orientation (information directionnelle). L'analyse du module nous permet de détecter les structures locales d'intérêt, également lorsque la position initiale de l'endroit d'analyse est proche du détail recherché.

Ainsi, l'idée proposée consiste à calculer un modèle d'apparence comme présenté dans la section 2, mais en remplaçant le vecteur de texture **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 6 :

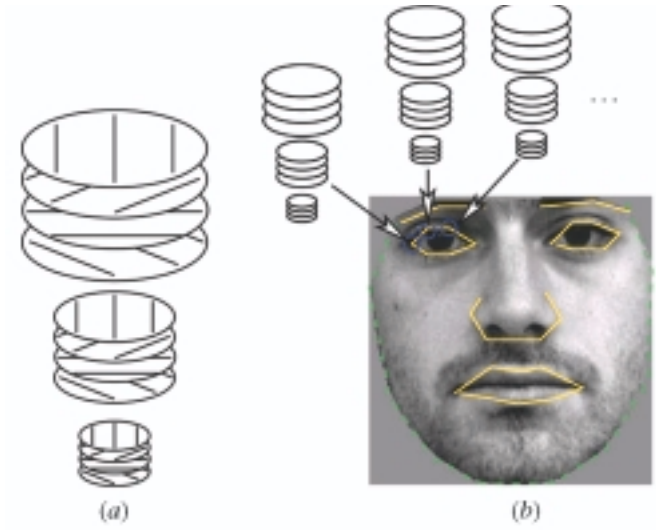


*Figure 6. (a) Représentation schématique d'un banc de filtres, ici de 3 échelles × 4 orientations (dans nos expériences, nous emploierons des bancs de $4 \times 8 = 32$ filtres). (b) Les réponses aux filtres sont calculées en chaque point annoté intérieur au visage, sur la texture de niveaux de gris ramenée à la forme moyenne des visages de l'ensemble d'apprentissage*

1. Les 32 réponses ne sont pas calculées en chaque pixel, mais uniquement aux points annotés, composant le vecteur **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. Ces quelques points sont repérés sur les visages d'apprentissage par une même personne, par identification de variations spatiales caractéristiques de la luminance des pixels de l'image, selon des directions déterminées (coins et bords des lèvres, des yeux et du nez).
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.

### 3.1 Adaptation du modèle

Nous proposons ici de vérifier l'intérêt, pour le problème de la détection automatique des traits faciaux, du modèle d'apparence hiérarchique par rapport à celui du modèle d'apparence standard décrit dans la section 2. Les deux modèles sont calculés à partir d'un ensemble d'apprentissage constitué de 37 images [48], dont un exemple est illustré sur la figure 7.a.

Le modèle hiérarchique basé sur des filtres de Gabor se montre expérimentalement plus robuste que le modèle standard vis-à-vis des conditions d'illumination. Les figures 7.c et 7.d montrent respectivement les résultats de l'adaptation des modèles standard et hiérarchique sur un visage cible inconnu, filmé sous des conditions d'éclairage très différentes de celles utilisées pour construire la base d'apprentissage.

Lorsque l'éclairage des séquences d'apprentissage et de test sont semblables, le modèle hiérarchique se montre toutefois moins précis que le modèle standard. On constate expérimentalement une erreur de détection moyenne de 2 pixels au niveau des traits faciaux, contre 1,6 pixels pour le modèle standard, à partir d'un ensemble de plusieurs visages tests et de différentes initialisations des modèles (par perturbations de leur pose). Cette relative imprécision est essentiellement due au fait que seul le module des réponses aux filtres de Gabor est pris en compte, et que l'amplitude de ce dernier varie lentement au voisinage de sa valeur maximale. La précision de la détection serait augmentée si l'on tenait compte de la phase locale de l'image filtrée [40] (au détriment dans ce cas de la robustesse du détecteur).

Le modèle hiérarchique se révèle quant à lui efficace lorsqu'il est initialisé loin du visage cible. La figure 8 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. …tant donné que seuls les points annotés internes au visage


*Figure 7. (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. (c) Détection par l'*AAM *standard. (d) Détection par l'*AAM *hiérarchique (*AAM *+ Gabor)*

sont pris en compte pour le calcul du modèle hiérarchique, la robustesse de celui-ci pourrait être supposée réduite ; les résultats montrent que la faible résolution spatiale du modèle hiérarchique (en terme de nombre 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.

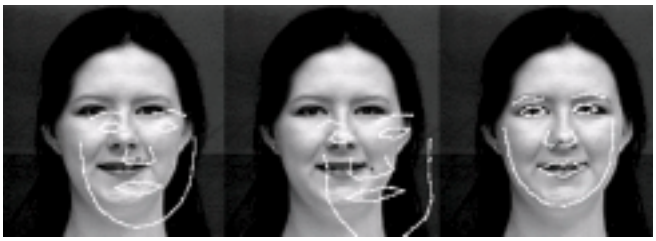

*Figure 3. De gauche à droite : Sur une nouvelle image inconnue, initialisation de la détection avec la forme moyenne* $\bar{\mathbf{s}}$*, positionnée avec une erreur de 30 pixels vers la droite et 30 pixels vers le bas ; Détection (divergente) obtenue avec l'*AAM *classique ; Détection obtenue avec l'*AAM *hiérarchique*

Nous observons également que la précision du modèle d'apparence hiérarchique pour la détection des traits de visages reste satisfaisante 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 a été faite en ne prenant qu'un nombre réduit de 20 puis de 10 images d'apprentissage (au lieu de 37) pour construire les modèles : une erreur de détection moyenne de 2 pixels au niveau des traits faciaux est préservée dans le cas du modèles hiérarchique. L'adaptation du modèle standard diverge dans les deux cas.

# 4. Reconnaissance d'expressions faciales

Nous proposons dans cette section une solution permettant de reconnaître l'expression faciale d'un visage cible, par discrimination linéaire, en partant du vecteur de paramètres d'apparence $\mathbf{c}_{op}$ qui lui est associé. Une analyse discriminante linéaire [17] est utilisée afin d'évaluer la qualité de la séparation des classes. L'apprentissage prédictif est effectué sur un ensemble constitué de sept classes correspondant aux six expressions de base en plus de l'expression neutre. Chaque classe contient 26 vecteurs $\mathbf{c}_i$ de taille 120 chacun pour la représentation standard et 170 chacun pour la représentation par *AAM* modifié. Le critère de similarité utilisé pour la classification d'un vecteur de paramètres $\mathbf{c}_{op}$ repose sur la distance de Mahalanobis comme montré dans l'équation (7), où $\mathbf{c}_{op}^{lda}$ est la projection du vecteur $\mathbf{c}_{op}$ dans l'espace discriminant, $\overline{\mathbf{c}}_j^{lda}$ est le vecteur moyenne de la classe $j$ dans l'espace discriminant et $\Sigma$ est la matrice de covariance des vecteurs $\mathbf{c}_{op}$ de chaque classe d'expression (on suppose ici l'égalité des sept matrices de covariance ($\Sigma = \Sigma_j \; \forall j$). Finalement $\mathbf{c}_{op}^{lda}$ sera assigné à la classe ayant la moyenne la plus proche.

$$d_M^{lda}(\mathbf{c}_{op}^{lda}, \overline{\mathbf{c}}_j^{lda}) = (\mathbf{c}_{op}^{lda} - \overline{\mathbf{c}}_j^{lda})^t \, \Sigma^{-1} (\mathbf{c}_{op}^{lda} - \overline{\mathbf{c}}_j^{lda}) \qquad (7)$$

Les résultats de reconnaissance à partir des modèles *AAM* standard et *AAM* modifié (calculé à l'aide d'une seule ACP) sont montrés dans les tableau 1 et 2 en considérant des vecteurs d'apparence tronqués, respectivement de dimension 60 et 50 (on vérifie expérimentalement qu'un nombre plus important de modes conduirait à des taux de bonne reconnaissance inférieurs).

Le tableau 3 illustre les capacités de reconnaissance de trente sujets humains entraînés, à partir du même ensemble test de 166 visages expressifs inconnus. Les sujets humains ont préalablement observés des visages expressifs de la base CMU.

L'analyse discriminante linéaire a pour but d'évaluer la séparabilité des classes comme illustré sur la figure 9. Une mesure de la séparation des classes dans le nouvel espace est donnée par le critère de Fisher : $trace(\mathbf{S}_w^{-1}\mathbf{S}_b)$, où $\mathbf{S}_w$ et $\mathbf{S}_b$ représentent respectivement les matrices de variance intra- et inter-classes dans l'espace de Fisher.

*Tableau 1. Matrice de confusion des expressions faciales, obtenue à partir du modèle d'apparence standard et de 166 images de test inconnues. Le pourcentage global de bonne classification est de 84,34 %. Colonne de droite : pourcentages de bonne classification*

| | neut. | col. | dég. | peu. | joi | sur. | tri. | % |
|---|---|---|---|---|---|---|---|---|
| neut. | 38 | 1 | 2 | 0 | 0 | 0 | 4 | 84,44 |
| col. | 3 | 13 | 0 | 0 | 0 | 0 | 1 | 76,47 |
| dég. | 0 | 0 | 18 | 0 | 1 | 0 | 1 | 90 |
| peu. | 1 | 0 | 0 | 14 | 3 | 0 | 0 | 82,35 |
| joi. | 0 | 0 | 0 | 4 | 22 | 0 | 0 | 84,62 |
| sur. | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 100 |
| tri. | 3 | 2 | 1 | 0 | 0 | 0 | 12 | 66,67 |

*Tableau 2. Matrice de confusion des expressions faciales, obtenue à partir du modèle* AAM *modifié et de 166 images de test inconnues. Le pourcentage global de bonne classification est de 83,73 %. Colonne de droite : pourcentages de bonne classification*

| | neut. | col. | dég. | peu. | joi | sur. | tri. | % |
|---|---|---|---|---|---|---|---|---|
| neut. | 33 | 3 | 2 | 0 | 0 | 0 | 7 | 73,33 |
| col. | 4 | 11 | 0 | 0 | 0 | 0 | 2 | 64,71 |
| dég. | 0 | 0 | 19 | 0 | 1 | 0 | 0 | 95 |
| peu. | 0 | 0 | 0 | 15 | 2 | 0 | 0 | 88,24 |
| joi. | 0 | 0 | 0 | 2 | 24 | 0 | 0 | 92,31 |
| sur. | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 100 |
| tri. | 1 | 1 | 1 | 0 | 0 | 1 | 14 | 77,78 |

*Tableau 3. Sommation de 30 matrices de confusion, obtenues par 30 sujets humains entraînés sur 166 images de test inconnues. Le pourcentage global de bonne classification est de 79,36 %. Colonne de droite : pourcentages de bonne classification.*

| | neut. | col. | dég. | peu. | joi | sur. | tri. | % |
|---|---|---|---|---|---|---|---|---|
| neut. | 1062 | 128 | 37 | 17 | 14 | 3 | 89 | 78,67 |
| col. | 19 | 397 | 37 | 4 | 0 | 0 | 53 | 77,84 |
| dég. | 16 | 123 | 441 | 7 | 1 | 4 | 8 | 73,5 |
| peu. | 26 | 12 | 59 | 313 | 46 | 35 | 19 | 61,37 |
| joi. | 10 | 1 | 11 | 18 | 720 | 15 | 5 | 92,31 |
| sur. | 0 | 0 | 1 | 33 | 1 | 654 | 1 | 94,78 |
| tri. | 74 | 62 | 29 | 7 | 0 | 3 | 365 | 67,59 |

# 5.Synthèse d'expressions faciales

Dans la première partie de cette section, nous proposons une modélisation des expressions faciales par régression linéaire entre les paramètres du modèle actif d'apparence standard [13]

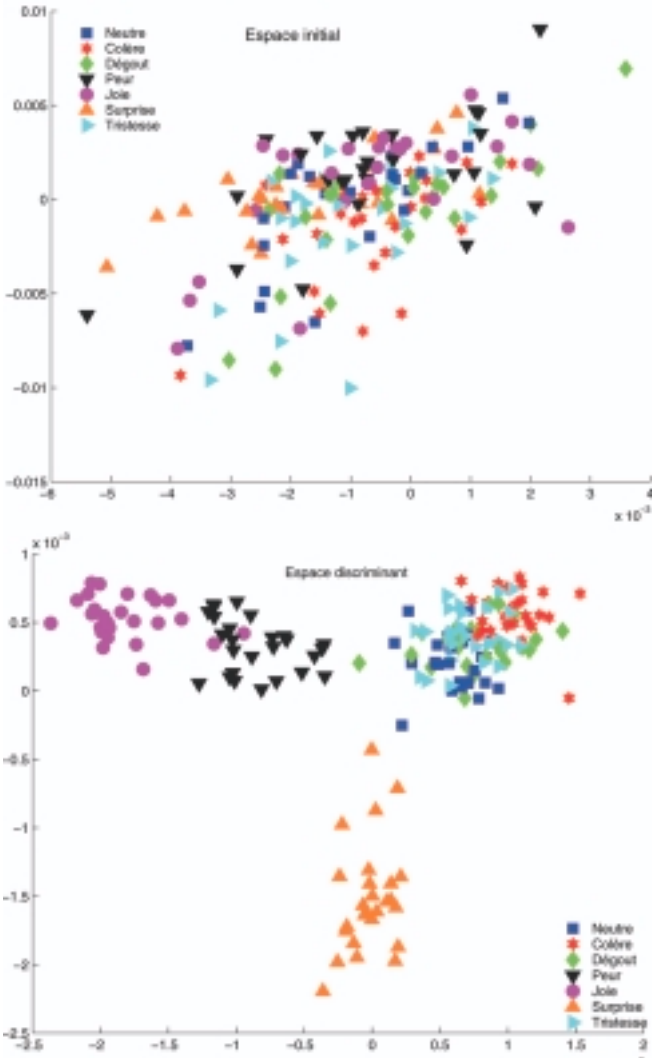


*Figure 9. Projections de l'ensemble d'apprentissage (composé de six classes d'expressions) sur, en haut, les duex axes principaux du modèle d'apparence, et en bas, les deux axes discriminants obtenus par analyse discriminante linéaire des vecteurs d'apparence de l'AAM standard (trace* $(S_w^{-1}S_b) = 37{,}9$*.) Les classes (les six sous-nuages de points) apparaissent dans ce dernier cas plus séparées*

et l'intensité de l'expression faciale (neutre, modérée ou intense). Le modèle linéaire direct ainsi obtenu est utilisé par Kang *et al.* [32] dans le cadre du filtrage de l'expression de visages. Nous proposons une extension de cette méthode à une application en synthèse d'expressions faciales. Notons dès à présent que cette modélisation linéaire ne nous permettra pas de synthétiser les mouvements fins et non linéaires des traits faciaux tels que ceux codés sous la forme d'une combinaison d'unités d'actions faciales définies par le système *FACS* (ce type de mouvement peut par exemple correspondre à une translation horizontale des commissures des lèvres suivie d'une translation verticale dans le cas de certaines expressions).

Dans une deuxième partie, nous proposons une modélisation des expressions faciales par régression linéaire couplant l'évolu-

tion des paramètres d'apparence (et non pas les paramètres eux-mêmes) à l'intensité modérée ou intense de l'expression faciale. Le modèle linéaire évolutif ainsi obtenu est utilisé dans le cadre de l'annulation et de la synthèse d'expressions faciales sur un visage donné. Enfin, des résultats expérimentaux illustrant les performances de ce modèle évolutif pour la synthèse d'expressions d'intensité variable seront donnés.

### 5.1 Modélisation directe de l'expression

En s'inspirant du modèle linéaire proposé par Cootes *et al.* pour représenter les paramètres d'apparence en fonction de l'orientation, un modèle linéaire reliant les paramètres d'apparence à l'intensité de l'expression faciale affichée peut s'écrire comme montré dans l'équation (8) :

$$\mathbf{c_e}(\mathcal{J}) = \mathbf{a_{e0}} + \mathbf{a}_{bfe1}\mathcal{J} + \varepsilon,$$

où $\mathcal{J}$ est un scalaire variant de $\mathcal{J} = 0$ pour indiquer une expression neutre à $\mathcal{J} = 1$ pour indiquer une expression intense, $\mathbf{a_{e0}}$ et $\mathbf{a_{e1}}$ sont des vecteurs de coefficients associés à l'expression faciale **e** (**e** = colère, dégoût, peur, joie, surprise ou tristesse), et $\varepsilon$ correspond à l'erreur de régression.
Pour chaque expression faciale **e**, l'apprentissage des vecteurs $\mathbf{a_{e0}}$ et $\mathbf{a_{e1}}$ se fait par régression linéaire sur un ensemble d'apprentissage en résolvant le système suivant :

$$\mathbf{Y} = \mathbf{X}\begin{pmatrix}\mathbf{a}_{\mathbf{e}0}^T \\ \mathbf{a}_{\mathbf{e}1}^T\end{pmatrix}. \tag{9}$$

$\mathbf{Y}$ est une matrice contenant dans les $m$ premières lignes, les vecteurs $\{\boldsymbol{c_i^T}\}$ correspondants aux visages de l'ensemble d'apprentissage affichant intensément l'expression faciale considérée et dans les $m$ lignes suivantes les vecteurs $\boldsymbol{c_i^T}$ correspondants aux visages de l'ensemble d'apprentissage affichant la même expression faciale, mais d'intensité moyenne. Les $m$ dernières lignes de $\mathbf{Y}$ contiennent les vecteurs $\boldsymbol{c_i^T}$ correspondants aux visages neutres de l'ensemble d'apprentissage. La deuxième colonne de $\mathbf{X}$ contient la valeur 1 dans les $m$ premières lignes, 0,5 dans les $m$ lignes suivantes et 0 dans les dernières lignes. La première colonne de $\mathbf{X}$ contient la valeur 1.
La solution par régression linéaire est donnée par :

$$\begin{pmatrix}\boldsymbol{a_{e0}^T} \\ \boldsymbol{a_{e1}^T}\end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{10}$$

#### 5.1.1. Filtrage de l'expression

Kang *et al.* [32] appliquent le modèle linéaire ainsi trouvé pour ramener l'expression faciale d'un visage cible inconnu vers une expression neutre. En effet, en connaissant *a priori* l'expression faciale affichée, les vecteurs $\mathbf{c_{e0}}$ and $\mathbf{c_{e1}}$ sont connus. La première étape consiste à trouver les paramètres $\mathbf{c}_{op}$ permettant d'adapter le modèle au visage cible. Le vecteur $\mathbf{c}_{op}$ nous permet ensuite d'estimer l'intensité de l'expression faciale en inversant le modèle défini dans l'équation (8).

$$\mathcal{J}_{est} = \mathbf{c}_{\mathbf{e}1}^{+}(\mathbf{c}_{op} - \mathbf{c_{e0}}), \tag{11}$$

où $\mathbf{c}_{\mathbf{e}1}^{+}$ est le pseudo-inverse de $\mathbf{c_{e1}}$. Un vecteur résiduel $\mathbf{c}_{res}$ contenant l'information relative à l'identité de la personne cible est obtenu en filtrant l'information contenue dans le modèle d'expression :

$$\mathbf{c}_{res} = \mathbf{c}_{op} - \mathbf{c_e}(\mathcal{J}_{est}). \tag{12}$$

En modifiant la valeur de $\mathcal{J}$ dans le vecteur $\mathbf{c_e}(\mathcal{J}_{est})$, il est possible de modifier l'intensité de l'expression. En particulier en imposant $\mathcal{J} = 0$ l'expression faciale affichée est entièrement filtrée.

$$\mathbf{c_e}(0) = \mathbf{c_{e0}} + \mathbf{c_{e1}} \times 0 = \mathbf{c_{e0}}. \tag{13}$$

Finalement, en conservant intact le vecteur $\mathbf{c}_{res}$, et en filtrant l'expression faciale ($\mathcal{J} = 0$), il est possible d'annuler l'expression affichée sur le visage cible inconnu et de la ramener vers une expression neutre comme montré sur la figure 10, à l'aide de l'équation suivante :

$$\mathbf{c}_{neutre} = \mathbf{c_e}(0) + \mathbf{c}_{res}. \tag{14}$$

#### 5.1.2 Synthèse d'expressions

La méthode proposée permet d'annuler l'expression faciale sur un visage expressif inconnu. En étendant cette même méthode à la synthèse, il est possible de générer une expression faciale sur un visage neutre inconnu, notamment sur un visage neutre synthétisé par annulation de son expression.
En effet, soit **e** l'expression désirée et soient $\mathbf{a_{e'0}}$ et $\mathbf{a_{e'1}}$ les paramètres correspondants. Il est alors possible d'estimer l'intensité de l'expression désirée sur le visage neutre de test.

$$\mathcal{J}'_{est} = \mathbf{a}_{\mathbf{e}1}^{+}(\mathbf{c}_{neutre} - \mathbf{a_{e0}}). \tag{15}$$


*Figure 10. a : visage joyeux inconnu. b : adaptation du modèle. c : annulation de l'expression par le modèle direct*

Le vecteur expressif moyen estimé à cette intensité est donné par :

$$\mathbf{c_e}(\mathcal{J}'_{est}) = \mathbf{a_{e0}} + \mathbf{a_{e1}}\mathcal{J}'_{est}. \quad (16)$$

Le nouveau résidu $\mathbf{c}_{res}$ est donné par :

$$\mathbf{c}_{res} = \mathbf{c}_{neutre} - \mathbf{c_e}(\mathcal{J}'_{est}). \quad (17)$$

L'intensité de l'expression faciale représentée par le vecteur $\mathbf{c_e}(\mathcal{J}'_{est})$ peut être contrôlée à l'aide du paramètre $\mathcal{J}$ dans l'équation (8). En particulier, il est possible de générer une expression intense en posant $\mathcal{J} = 1$. Il est alors possible de synthétiser une expression intense comme montré sur la figure 11, à l'aide de l'équation suivante :

$$\mathbf{c}_{intense} = \mathbf{c_e}(1) + \mathbf{c}_{res}. \quad (18)$$

### 5.2 Un modèle linéaire évolutif

Le modèle linéaire évolutif proposé consiste à déterminer pour chaque expression faciale **e** les paramètres de l'équation :

$$\boldsymbol{c_{op}} - \boldsymbol{c_{op_n}} = \boldsymbol{a_{e1}}\mathcal{J} + \varepsilon, \quad (19)$$

$\boldsymbol{c_{op}} - \boldsymbol{c_{op_n}}$ étant la différence entre les paramètres d'apparence optimaux $\boldsymbol{c_{op}}$ permettant de synthétiser un visage ressemblant au visage cible, les vecteurs $\boldsymbol{c_{op_n}}$ permettant de synthétiser ce même visage avec une expression neutre, et $\varepsilon$ correspondant à l'erreur de régression. Comme dans l'équation (8), $\mathcal{J}$ est un scalaire variant de 0 à 1. Le vecteur $\mathbf{c_{e1}}$ dépend de l'expression faciale affichée et est appris par régression linéaire sur un ensemble d'apprentissage initial en résolvant le système suivant :

$$\mathbf{Z} = \mathbf{X}\boldsymbol{a}_{\boldsymbol{e1}}^{T} + \varepsilon, \quad (20)$$

où $\mathbf{Z}$ est constituée par les $2m$ premières lignes de la matrice $\mathbf{Y}$ décrite en 5.1 dont on a soustrait les paramètres codant les mêmes visages avec une expression neutre. La deuxième colonne de $\mathbf{X}$ contient la valeur 1 dans les $m$ premières lignes et 0,5 dans les $m$ lignes suivantes. La première colonne de $\mathbf{X}$ est composée de valeurs 1.

Il est ainsi possible d'annuler l'expression d'un visage inconnu, représenté par le vecteur d'apparence $\boldsymbol{c_{op}}$, à l'aide de l'équation :

$$\boldsymbol{c_{op_n}} = \boldsymbol{c_{op}} - \boldsymbol{a_{e1}}\mathcal{J}, \quad \text{en supposant } \mathcal{J} = 1. \quad (21)$$

De la même manière, une nouvelle expression $\mathbf{e}'$ peut être synthétisée sur un visage neutre inconnu représenté par $\boldsymbol{c_{op_n}}$, à l'aide du vecteur $\boldsymbol{c_{op_{e'}}}$ donné par :

$$\boldsymbol{c_{op_{e'}}} = \boldsymbol{c_{op_n}} + \boldsymbol{a_{e'1}}\mathcal{J}. \quad (22)$$


*Figure 11. Génération de six expressions faciales synthétiques par le modèle linéaire direct en partant du visage neutre synthétique de la figure 10.c. a : colère. b : dégoût. c : peur. d : joie. e : surprise. f : tristesse*

Le modèle linéaire évolutif proposé en 5.2 constitue une approche plus directe et plus intuitive pour la synthèse d'expressions faciales. Dans cette approche, les hyperplans permettant de modéliser chacune des six expressions faciales de base passent tous par la même origine; alors que dans la méthode directe décrite en 5.1 chaque hyperplan caractérisant une expression faciale possède une origine propre dépendant de cette expression. En ce qui concerne la qualité visuelle des images synthétisées, les performances du modèle évolutif (figure 12) sont comparables sinon supérieures à celles du modèle classique (figure 11) : l'adaptation du modèle évolutif à la forme de chacune des sept expressions faciales est notamment plus précise. Ceci a été vérifié expérimentalement par mesure des erreurs de position des points de la forme des visages, en partant d'un ensemble de visages initialement neutres extraits de séquences de la base CMU, rendus artificiellement expressifs à l'aide des modèles directes et évolutifs, puis comparés aux visages expressifs connus extraits des mêmes séquences.

### 5.3 Évolution de l'expression sur une vidéo

Afin d'analyser le comportement temporel des modèles linéaires obtenus en 5.1 et 5.2, une série d'expériences sera effectuée sur un ensemble de 8 vidéos montrant une évolution graduelle de l'expression faciale allant d'une expression neutre à une expression intense pour différentes personnes.

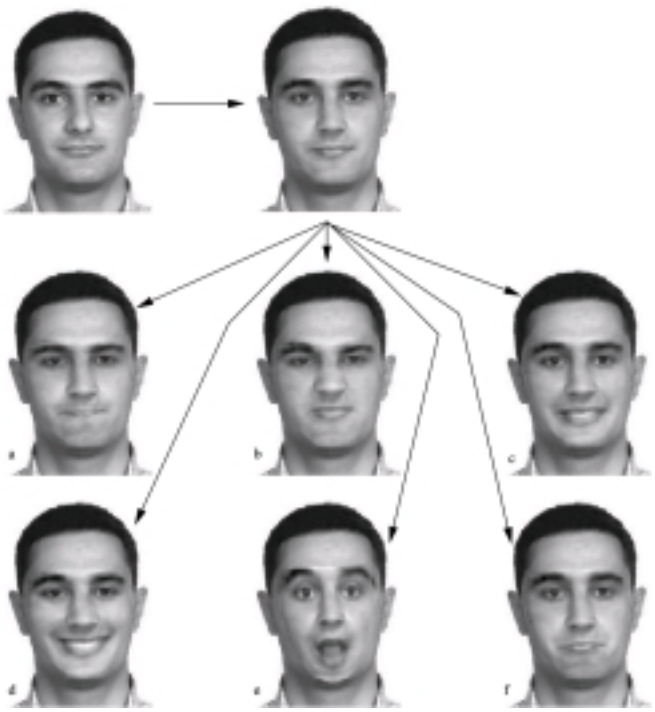

*Figure 12. Génération de six expressions faciales synthétiques à l'aide du modèle linéaire évolutif en partant d'un visage neutre. a : colère. b : dégoût. c : peur. d : joie. e : surprise. f : tristesse*

Le modèle actif d'apparence classique est adapté sur chaque image de la séquence vidéo, en initialisant les paramètres d'apparence et de pose d'après l'adaptation de l'image précédente comme montré sur la figure 13 (deuxième ligne).

Au niveau de chaque image de la séquence vidéo, le vecteur de paramètres $\mathbf{c}_{op}$ obtenu permet d'estimer l'intensité de l'expression affichée $\mathcal{J}_{est}$ à l'aide de l'équation (11) ainsi que le vecteur de paramètres d'apparence prédit à cette intensité $\mathbf{c}_{\mathbf{e}}(\mathcal{J}_{est})$ à l'aide de l'équation linéaire (8). De même, en supposant que la première image de la séquence correspond à un visage neutre, il est possible de calculer pour chaque image de la séquence l'évolution $\mathbf{c}_{op} - \mathbf{c}_{op_n}$ par rapport au visage neutre $\mathbf{c}_{op_n}$ initial.

Nous remarquons que pour chaque expression faciale, les paramètres $\mathbf{c}_{\mathbf{e}}(\mathcal{J}_{est})$ tendent à suivre une trajectoire temporelle bien définie pouvant être approximée par le modèle linéaire calculé en 5.1.

Ceci est illustré par la figure 14 montrant l'évolution du premier coefficient de $\mathbf{c}_{\mathbf{e}}(\mathcal{J}_{est})$ sur huit séquences d'images. Les séquences montrent des visages évoluant d'une expression neutre vers une expression intense de surprise. La ligne droite montre l'évolution de la droite $\mathbf{c}_{\mathbf{e}}(\mathcal{J})$, avec $\mathcal{J}$ variant linéairement de 0 à 1.

De même, les paramètres $\mathbf{c}_{op} - \mathbf{c}_{op_n}$ tendent à suivre une trajectoire temporelle bien définie pouvant être approximée par le modèle linéaire calculé en 5.2. Ceci est illustré par la figure 15 montrant l'évolution du premier coefficient de $\mathbf{c}_{op} - \mathbf{c}_{op_n}$. La ligne droite montre l'évolution temporelle de $\boldsymbol{a}_{\mathbf{e}1}\mathcal{J}$, avec $\mathcal{J}$ variant linéairement de 0 à 1.

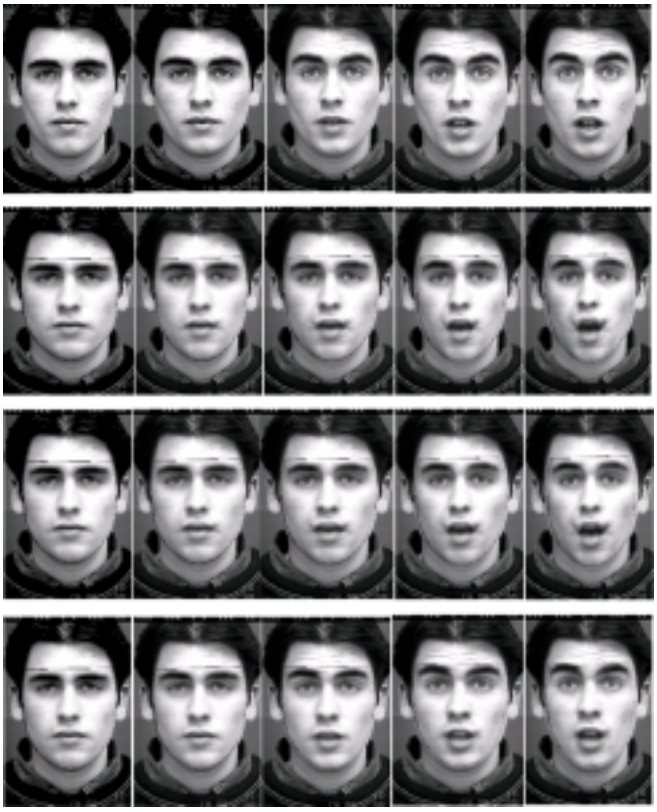

*Figure 13. Première ligne : visage initial surpris. Deuxième ligne : adaptation du modèle d'apparence. Le modèle actif d'apparence peut être utilisé pour effectuer un suivi de visages expressifs dans une séquence vidéo. Troisième ligne : visage surpris synthétique prédit à l'aide de l'équation linéaire (8). Dernière ligne : visage surpris synthétique prédit à l'aide de l'équation linéaire (22)*

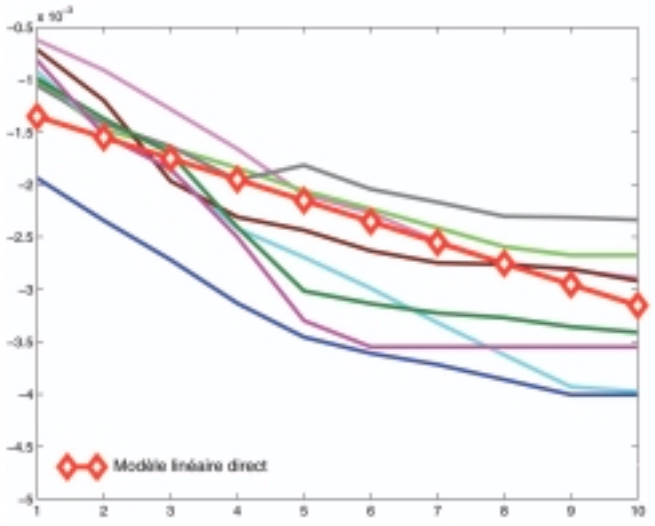


*Figure 14. Évolutions temporelles du premier mode de $\mathbf{c}_{\mathbf{e}}(\mathcal{J}_{est})$ associées à huit séquences d'images montrant des visages neutres devenant surpris (chaque séquence vidéo est composée de dix images). Ligne droite : évolution du premier mode du modèle linéaire **direct***

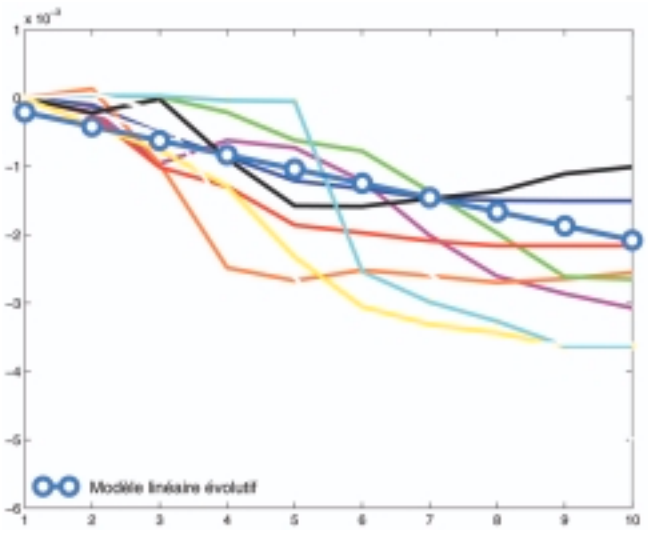


*Figure 15. Évolutions temporelles du premier mode de* $\mathbf{c}_{op} - \mathbf{c}_{op_n}$. *Ligne droite : évolution du premier mode du modèle linéaire* ***évolutif***

# 6. Conclusion

En partant du cadre formel des modèles actifs d'apparence de Cootes *et al.*[11], nous avons proposé dans cet article une méthode permettant de construire un modèle d'apparence hiérarchique, basé sur des filtres de Gabor. Ce premier travail se place dans la problématique de la détection robuste 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. Le modèle hiérarchique reposant sur une représentation du visage dans un espace multi-résolution (transformée de Gabor), il ne permet pas de synthétiser l'apparence de nouveaux visages. Il peut cependant être utile pour la reconnaissance des gestes faciaux (expressions faciales, mimiques, clignements des yeux, etc.). Celle-ci peut dans ce cas se faire à partir des vecteurs d'apparence calculés selon la méthode classique ou la méthode hiérarchique.

Dans une deuxième partie, nous avons montré l'intérêt de la représentation par le modèle actif d'apparence standard pour la reconnaissance d'expressions faciales. Deux approches de la modélisation d'expressions faciales par régression linéaire sur un ensemble d'apprentissage ont été abordées. Les modèles linéaires direct et évolutif ainsi obtenus ont été utilisés pour le filtrage et la synthèse d'expressions faciales. Le modèle évolutif proposé constitue une approche plus simple de la synthèse d'expressions faciales. En effet, dans ce cas, les hyperplans permettant de modéliser chacune des six expressions faciales de base passent tous par la même origine alors que pour le modèle direct chaque hyperplan caractérisant une expression faciale possède une origine propre. Les qualités visuelles des images synthétisées par ces deux modèles sont toutefois sensiblement comparables.

Nous étudions actuellement l'extension des approches proposées pour la synthèse de mélanges d'expressions. Nous étudions également l'utilisation de mesures de similarité robustes pour rendre l'adaptation du modèle *AAM* moins dépendante des occultations partielles et des variations d'illumination.

# Références


[1] M.S. BARTLETT, G. LITTLEWORT, B. Braathen, T.J. Sejnowski, J.R. Movellan, An approach to automatic analysis of spontaneous facial expressions, In *Proceedings of Advances in Neural Information Processing Systems*, 2002.

[2] M.S. BARTLETT, G. LITTLEWORT, I. FASEL, J. R. MOVELLAN, Face Detection, Facial Expression Recognition: Development and Applications to Human Computer Interaction, In *IEEE workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, Madison, U.S.A., June, 2003.

[3] D. BEYMER, Vectorizing face images by interleaving shape and texture computations, MIT, U.S.A., A.I. Memo n° 1537, September 1995.

[4] M.J. BLACK, Y. YACOOB, Recognizing Facial Expressions in Image Sequences Using Local Parametrized Models of Image Motion, *International Journal of Computer Vision*, 25 (1), p. 23–48, 1997.

[5] V. BLANZ, T. VETTER, Face Recognition based on fitting a 3D morphable model, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), September, 2003.

[6] E. BOYLE, A.H. ANDERSON, A. NEWLANDS, The Effects of Visibility on Dialogue in a Cooperative Problem Solving Task, *Language and Speech*, 37, p. 1–20, 1994.

[7] C. BREGLER, M. COVEL, M. SLANEY, Video Rewrite: Driving Visual Speech with Audio, In *ACM Siggraph*, p. 353–360, 1997.

[8] R. BRUYER, *Le visage et l'expression faciale : approche neuropsychologique*, Pierre Mardaga éditeur. Collection Psychologie et Sciences Humaines, n° 118, 1983.

[9] E.S. CHUANG, H. DESHPANDE, C. BREGLER, Facial Expression Space Learning, In *IEEE Pacific conference on computer graphics and applications*, october 2002.

[10] J. COHN, A. ZLOCHOWER, J.J. LIEN, T. KANADE, Feature-point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression, In *International Conference on Automatic Face and Gesture Recognition*, p. 396–401, Nara, Japan, 1998.

[11] T.F. COOTES, G.J. EDWARDS, C.J. TAYLOR, Active Appearance Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), p. 681–685, June 2001.

[12] T.F. COOTES, P. KITTIPANYA-NGAM, Comparing variations on the active appearance model algorithm, In *British Machine Vision Conference*, p. 837–846, Cardiff University, September 2002.

[13] T.F. COOTES, K. WALKER, C.J. TAYLOR, View-Based Active Appearance Models, In *International Conference on Automatic Face and Gesture Recognition*, p. 227–232, Grenoble, France, March 2000.

[14] G. DONATO, M.S. BARTLETT, J.C. HAGER, P. EKMAN, T.J. SEJNOWSKI, Classifying Facial Actions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10), p. 974–988, October 1999.

[15] B. DRAPER, K. BAEK, M.S. BARTLETT, R. BEVERIDGE, Recognizing Faces with PCA and ICA, *Computer Vision and Image Understanding*, 91(1/2), p. 115–137, July 2003

[16] I.L. DRYDEN, K.V. MARDIA, *Statistical Shape Analysis*, John Wiley, 1998

[17] S. DUBUISSON, F. DAVOINE, M. MASSON, A solution for facial expression representation and recognition, *Signal Processing: Image Communication*, 17(9), p. 657–673, October 2002.

[18] G.J. EDWARDS, T.F. COOTES, C.J. TAYLOR, Face Recognition Using Active Appearance Models, In *European Conference of Computer Vision*, p. 581–695, 1998.

[19] P. EKMAN, *Emotion in the human face*, Cambridge University Press, 1982.

[20] P. EKMAN, W. FRIESEN, Facial Action Coding System: A Technique for the Measurement of Facial Movement, *Palo Alto, Calif.: Consulting psychologists press*, 1978.

[21] I. ESSA, A. PENTLAND, Coding, Analysis, Interpretation, Recognition of Facial Expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), p. 757–763, 1997.

[22] T. EZZAT, G. GEIGER, T. POGGIO, Trainable Videorealistic Speech Animation, In *ACM Siggraph* July, San Antonio, Texas, 2002.

[23] T. EZZAT, T. POGGIO, Facial analysis and synthesis using image-based models, In *International Conference on Automatic Face and Gesture Recognition*, p. 116–121, 1996.

[24] S.B. GOKTURK, J.-Y. BOUGUET, R. GRZESZCZUK, A data-driven model for monocular face tracking, In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[25] H. GUPTA, A.K. AGRAWAL, T. PRUTHI, C. SHEKHAR, R. CHELLAPPA, An experimental evaluation of linear and kernel-based methods for face recognition, In *International workshop on applications of computer vision*, Orlando, Floride, December 2002.

[26] B. HEISELE, P. HO, T. POGGIO, Face Recognition with Support Vector Machines: Global Versus Component-based Approach, In *International Conference on Computer Vision*, Vancouver, Canada, p. 688–694, July 2001.

[27] E. HJELMAS, B. LOW, Face detection: a survey, *Computer Vision and Image Understanding*, 83, p. 235–274, 2001.

[28] H. HONG, H. NEVEN, C. VON DER MALSBURG, Online Facial Expression Recognition based on Personalized Gallery, In *Intl. Conference on Automatic Face and Gesture Recognition*, p. 354–359, Nara, Japan, 1998

[29] X. HOU, S. LI, H. ZHANG, Q. CHENG, Direct appearance models, In *Intl. Conference on Computer Vision and Pattern Reocognition*, p. 828–833, 2001.

[30] C. HU, R. FERIS, M. TURK, Active wevelet networks for face alignment, In *British machine vision conference*, East Eaglia, Norwich, U.K., 2003.

[31] T. KANADE, J. COHN, Y.L. TIAN, Comprehensive database for facial expression analysis, In *International Conference on Automatic Face and Gesture Recognition*, p. 46–53, Grenoble, France, March 2000.

[32] H. KANG, T.F. COOTES, C.J. TAYLOR, Face Expression Detection and Synthesis using Statistical Models of Appearance, In *Measuring Behavior*, p. 126–128, Amsterdam, The Netherlands, August 2002.

[33] A. LANITIS, C.J. TAYLOR, T.F. COOTES, Automatic interpretation and coding of face images using flexible models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), p. 743–756, July 1997.

[34] C. LIU, Gabor-based Kernel PCA with Fractional Power Polynomial Models for Face Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5), p. 572–581, May 2004.

[35] M.J. LYONS, J. BUDYNEK, S. AKAMATSU, Automatic Classification of Single Facial Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12), p. 1 357–1 362, December 1999

[36] 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.

[37] I. MATTHEWS, S. BAKER, Active Appearance Models Revisited, Technical Report CMU-RI-TR-03-02 The Robotics Institute, Carnegie Mellon University, April 2003.

[38] A. MEHRABIAN, Communication without Words, *Psychology Today*, 2(4), p. 53–56, 1968.

[39] B. MOGHADDAM, Principal manifolds and probabilistic subspaces for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6), June 2002.

[40] A. NICOULIN, *Analyse d'images par spectre local de phase*, Presses Polytecniques et Universitaires Romandes, Collection META, 1990.

[41] C. PADGETT, G. COTTRELL, Representing face images for emotion classification, In *Advances in Neural Information Processing Systems, MIT Press*, volume 9, p. 894–900, Cambridge, MA., 1997.

[42] C. PADGETT, G. COTTRELL, R. ADOLPHS, Categorical Perception in Facial Emotion Classification, In *ACM Siggraph*, p. 75–84, 1996.

[43] I. PANDZIC, R. FORCHHEIMER, MPEG-4 *Facial Animation - The standard, implementations, applications*, John Wiley, 2002.

[44] M. PANTIC, L.J.M. ROTHKRANTZ, Automatic Analysis of Facial Expressions: The State of the Art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), p. 1 424–1 445, December 2000

[45] P. PENEV, J. ATICK, Local Feature Analysis: A general Statistical Theory for Object Representation, *Network: computation in neural systems*, 7(3), p. 477–500, 1996.

[46] F. PIGHIN, J. HECKER, D. LISCHINSKI, R. SZELISKI, D.H. SALESIN, Synthesizing Realistic Facial Expressions from Photographs, In *ACM International conference on computer graphics and interactive techniques*, 1998.

[47] B. SCHOLKOPF, A. SMOLA, K.-R. MULLER, Kernel principal component analysis, In *Proc. of Artificial Neural Networks – ICANN*, Berlin, 1997.

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

[49] Y.-L. TIAN, T. KANADE, J.F. COHN, Recognizing action units for facial expression analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), p. 97–115, February 2001.

[50] T. VETTER, N.F. TROJE, Face Recognition based on fitting a 3D morphable model, *Journal of the optical society of america A*, 14(9), p. 2 152–2 161, 1997.

[51] P. VIOLA, M. JONES, Robust real-time object detection, In *International workshop on statistical computational theories of vision – modeling, learning, computing and sampling*, Cancouver, Canada, July 2001.

[52] Laurenz WISKOTT, Jean-Marc FELLOUS, Norbert KRÜGER, Christoph VON DER MALSBURG, Face Recognition by Elastic Bunch Graph Matching, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, In L. C. Jain, U. Halici, I. Hayashi, S. B. Lee, CRC Press, 0-8493-2055-0, 11, p. 355–396 ,1999.

[53] M. H. YANG, Face Recognition Using Kernel Methods, In *Advances in Neural Information Processing Systems*, 14, p. 215–220, 2002.

[54] M.-H. YANG, D. KRIEGMAN, N. AHUJA, Detecting faces in images: a survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), p. 34–58, January 2002.

[55] W. ZHAO, R. CHELLAPPA, A. ROSENFELD, P.J. PHILLIPS, Face recognition: A literature survey, CVL, University of Maryland, U.S.A., October 2000.

Franck **Davoine**

Docteur INPG (décembre 1995) en Signal, Image, Parole, Franck Davoine est actuellement Chargé de recherche au CNRS, au sein de l'UMR Heudiasyc (Université de Technologie de Compiègne / CNRS). Ses domaines d'intérêt incluent l'analyse et le suivi de visages, et plus généralement la reconnaissance du comportement facial.

Van Mô **Dang**

Docteur UTC (décembre 1998) en Contrôle des Systèmes, Van Mô Dang est Maître de conférences à l'Université de Technologie de Compiègne, au sein de l'UMR Heudiasyc. Ses domaines d'intérêt incluent l'analyse d'images et la reconnaissance statistique de formes. Il est mis en disponibilité depuis le 1 septembre 2004.

Bouchr **Abboud**

Doctorante UTC depuis octobre 2001, Bouchra Abboud effectue ses recherches au sein de l'UMR Heudiasyc. Ses domaines d'intérêt portent sur la modélisation des visages et de leurs expressions.