Une solution pour la reconnaissance d'expressions de visages

Séverine DUBUISSON, Franck DAVOINE, Mylène MASSON

Laboratoire Heudiasyc - Université de Technologie de Compiègne
BP 20529. 60205, Compiègne Cedex, France
sdubuiss@hds.utc.fr

Résumé –
La procédure de représentation et de classification pour la reconnaissance d'expressions que nous proposons considère la partie interne de visages obtenue par une détection manuelle de traits caractéristiques (yeux, nez et bouche). Une Analyse en Composantes Principales nous permet de représenter ces visages dans un espace de dimension réduite, défini par des bases adaptées à l'ensemble d'apprentissage d'expressions. Nous montrons comment le fait de sélectionner la meilleure base de projection permet une meilleure discrimination linéaire pour la reconnaissance d'expressions de visages. Les résultats prouvent enfin la robustesse de cette méthode de reconnaissance, comparée aux méthodes habituelles de projection. Enfin, des tests comparatifs permettent de dire quelles caractéristiques du visage et quelles représentations semblent être plus adaptées à la reconnaissance d'expressions.

Abstract –
The representation and classification process we propose for the facial expression recognition problem considers the internal part of faces given by a manual facial feature detection (eyes, nose and mouth). PCA gives a lower dimensional representation space for the faces, defined by basis adapted to the class recognition problem. We show how to select the best projection basis and the improvement of the recognition in this basis. Results prove the robustness of this method, if we compare with usual projecting and classification algorithms. Finally, comparative test inform of which features and representation of faces seem to be more adapted to the facial expression recognition problem.

1 Introduction

L'analyse de visages est une discipline en traitement d'images qui se décompose en deux phases : l'extraction de caractéristiques et la reconnaissance (de visage, de genre, de posture, de race, d'expression, etc.). D'une manière générale, une expression est une combinaison des Unités d'Action (AUs) définies par Ekman et Friesen [3] : elle se caractérise par un mouvement facial local provoqué par la contraction ou l'étirement des muscles du visage. Une expression peut ainsi être reconnue de deux manières : en travaillant sur une séquence vidéo et en analysant un mouvement sur le visage, ou bien sur des images statiques, où l'on recherche une forme et une texture du visage spécifiques (par des outils d'analyse statistique, de filtrage local ou encore la mesure de traits caractéristiques). Les chercheurs se sont d'abord intéressés à l'analyse du mouvement : Basili [1] a défini un modèle de mouvement global apparent sur le visage pour chacune des six expressions fondamentales (peur, étonnement, dégoût, colère, tristesse et joie). Yacoob et Davis [8] analysent le flot optique : ils encadrent manuellement des traits caractéristiques sur la première image puis effectuent un suivi de ceux-ci tout au long de la séquence. Ils peuvent ainsi reconnaître l'expression du visage en analysant son mouvement (i. e. sa déformation) interne. Pour une analyse statistique, on doit dans un premier temps apprendre ce que l'on recherche : Moghadam et Pentland [6] ont proposé une méthode d'apprentissage de l'apparence d'une expression faciale en utilisant une décomposition en vecteurs propres de l'espace image, puis reconnaissent ou non une expressions grâce à une mesure de vraisemblance. Les méthodes de filtrage local permettent de changer le domaine de définition d'une image, faisant ressortir des traits caractéristiques et donc aidant à leur détection : de nombreux chercheurs [5, 2] ont utilisé les ondelettes de Gabor afin de mettre en évidence certaines parties du visage, qui varient d'une expression à l'autre. Nous allons présenter, dans les sections qui suivent, notre méthode de reconnaissance d'expressions, qui utilise une approche statistique, après avoir extrait l'information importante de l'expression des visages.

2 Extraction de masques faciaux

L'expression d'un visage se manifeste de manière significative au niveau des traits caractéristiques (comme les yeux ou la bouche). Pour cette raison, leur extraction est une étape importante avant une analyse plus fine pour la reconnaissance d'expressions. Nous ne considérerons ainsi, dans la suite, que la partie interne d'un visage afin d'en reconnaître l'expression. Pour cela, nous sélectionnons manuellement 4 points caractéristiques (les centres des pupilles, le milieu de la bouche et le nez) afin d'effectuer une normalisation géométrique de manière à ce que ces 4 points soient positionnés à des coordonnées fixes. On applique deux transformations affines : une sur le haut du visage (au dessus du nez) afin de normaliser le triangle "yeux-nez" et une sur le bas du visage afin de normaliser le triangle "yeux-bouche". Cette normalisation géométrique est illustrée par la figure 1. Afin de compenser les problèmes de changement d'illumination, nous effectuons une normalisation en luminance des visages par spécification d'histo-

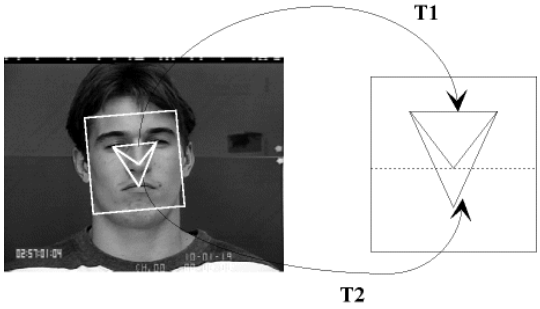

FIG. 1 – Principe de la normalisation manuelle : 4 points caractéristiques sont manuellement positionnés (les deux pupilles, la pointe du nez et le milieu des lèvres) donnant lieu à deux transformations affines T_1 et T_2 pour ramener ces points à des positions fixes sur l'image cible. La texture interne subit les mêmes transformations.

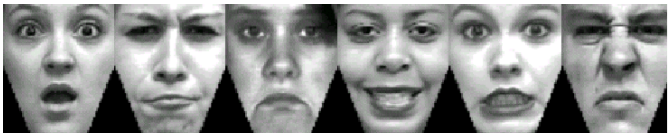

FIG. 2 – Masques extraits de visages de manière manuelle pour différentes expressions.

grammes : on utilise pour cela un histogramme référence (celui d'une image de la base dont les conditions d'illumination sont satisfaisantes). Enfin, on extrait un masque facial en utilisant les positions des points caractéristiques : la figure 2 montre le résultat final sur 6 visages soumis à différentes expressions. Dans la suite de l'article, la reconnaissance de l'expression d'un visage se fera uniquement sur ces masques. Chaque individu x_i (visage) est un vecteur de taille 4200 (60×70 pixels) appartenant à une des 6 classes d'expression.

3 Analyse d'expressions de visages

L'algorithme est divisé en trois parties principales :

1. Une Analyse en Composantes Principales (ACP), qui trouve un sous-espace dont la base de vecteurs correspond aux directions de variance maximum de l'espace original.
2. La recherche d'une base de projection optimale via une sélection des composantes principales les plus utiles pour le problème de reconnaissance d'expressions.
3. Une Analyse Discriminante Linéaire (ADL) afin d'obtenir le sous-espace le plus discriminant pour la classification.

L'ensemble d'apprentissage S est construit en utilisant N masques de visages (i. e. vecteurs) appartenant à six classes d'expression distinctes ($\frac{N}{6}$, ou N_c vecteurs par classe d'expression).

3.1 ACP et recherche de bases optimales

Afin de réduire la dimension de l'espace de travail, on effectue une ACP sur l'ensemble d'apprentissage S. Les valeurs propres λ_i de la matrice de covariance $C = SS^T$ sont rangées dans l'ordre décroissant et la qualité globale de représentation de l'ensemble initial est donnée par le taux d'inertie $r_{I_M} = \frac{\lambda_1 + \cdots + \lambda_M}{trace(C)}$. La méthode habituellement employée pour la classification est de projeter l'ensemble initial dans la base formée par les M premiers vecteurs propres, où M est le nombre de valeurs propres nécessaires à l'obtention d'un certain taux d'inertie. Cependant, nous ne sommes pas sûrs que les composantes principales liées aux plus fortes valeurs propres donnent des projections optimales pour résoudre un problème de reconnaissance d'expressions. Pour cette raison, on considère une procédure itérative de validation croisée dont le but est de minimiser une erreur de généralisation du système de classification utilisant le critère de Fisher ([4]) $\text{trace}(S_W^{-1} S_B)$, où S_B et S_W sont respectivement les matrices de variance inter- et intra-classes, définies par :

$$S_B = \sum_{i=1}^{c} (\tilde{m}^i - \tilde{m})(\tilde{m}^i - \tilde{m})^t$$
$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{N_c} (x_j^i - \tilde{m}^i)(x_j^i - \tilde{m}^i)^t$$

avec c le nombre de classes, $\tilde{m}^i$ la moyenne de la classe i, $\tilde{m}$ la moyenne totale, x_j^i l'individu j de la classe i et N_c le nombre d'individus par classe.

L'ACP a permis de déterminer N composantes principales parmi lesquelles nous recherchons les K meilleures en utilisant une méthode de sélection "pas à pas" dans laquelle on introduit progressivement les variables selon le principe général suivant :

- A l'étape 1, on cherche la meilleure composante principale parmi les N disponibles.
- A l'étape k, on cherche la composante principale (parmi les $N - k + 1$ restantes) qui, ajoutée à celles retenues lors des étapes précédentes, est la meilleure.

On classe ainsi toutes les composantes principales disponibles en sortie d'ACP. Une étape consiste à effectuer les opérations suivantes (cette sélection est répétée N_r fois, à partir d'ensembles d'apprentissage et de test différents échantillonnés dans l'ensemble de données initial - 80% pour l'apprentissage, 20% pour le test) :

1. Sélectionner une composante parmi les disponibles et l'ajouter à l'ensemble précédemment composé (initialement vide)
2. Construire le sous-espace de projection correspondant formé des vecteurs propres associés
3. Calculer les matrices de variance intra- (S_W) et inter-classes (S_B) de l'ensemble projeté
4. Calculer le critère de Fisher : $\text{trace}(S_W^{-1} S_B)$

Après avoir sélectionné toutes les composantes disponibles, nous gardons celle qui aura maximisé le critère de Fisher et l'ajoutons à celles sélectionnées aux étapes précédentes. L'algorithme détermine ainsi les composantes principales qui donnent le meilleur résultat en terme de classification : la meilleure paire de composantes est sélectionnée, puis le meilleur triplet, etc. Ce système est moyenné pour un certain nombre d'ensembles d'apprentissage et test (choisis aléatoirement). Cet algorithme nous permet de déterminer le nombre optimal K de composantes principales nécessaires (dans leur ordre d'importance) pour lequel l'erreur de classification généralisée est minimale : la figure 3 présente

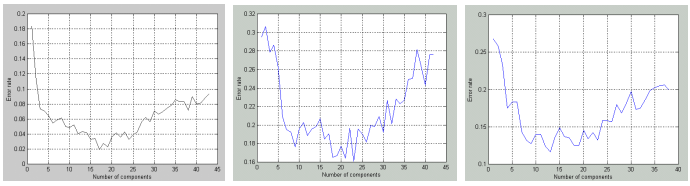

FIG. 3 – Profil de l'erreur de généralisation pour les classifieurs binaires (de gauche à droite) "Tristesse/joie", "Joie/Peur" et "Peur/Colère". Le minimum des courbes indique le nombre optimal K de composantes principales nécessaires pour une bonne classification ($K = 17, 27, 13$).

l'erreur de généralisation résultant de la moyenne des erreurs observées sur les N_r ensembles de test. Ces composantes correspondent aux vecteurs de l'ensemble d'apprentissage projetés dans la base optimale définie par les vecteurs propres associés.

3.2 ADL

Une fois que nous avons projeté l'ensemble dans la base optimale, on effectue une ADL sur celui-ci, afin de concentrer et séparer de manière optimale les classes qui y sont représentées. Cette approche est décrite, dans le cas d'un problème à deux classes, dans [5]. D'une manière générale, on cherche à trouver un sous-espace dans lequel les différentes classes définies par un ensemble d'individus sont représentées de manière compacte et séparées les unes des autres le mieux possible : on maximise la distinction entre les groupes tout en minimisant les variations au sein de ces groupes. Pour cela, on doit donc maximiser la variance inter-classes, tout en minimisant la variance intra-classes : on cherche les axes qui discriminent le mieux les classes (et non pas les individus). Ainsi, les axes factoriels sont les vecteurs propres associés aux plus grandes valeurs propres de la matrice $S_W^{-1} S_B$: on cherche à maximiser la trace de cette matrice. L'espace discriminant fourni par l'ADL est de dimension $(c-1)$, où c correspond au nombre de classes présentes dans l'ensemble d'apprentissage. Afin de déterminer à quelle classe appartient une nouvelle image, on la projette d'abord dans le sous-espace optimal (obtenu après sélection des meilleures composantes), puis dans le sous-espace de séparation des classes, formé par les axes factoriels, et on calcule sa distance (de Mahalanobis) à chacune des classes de cet ensemble : on l'associe à celle dont elle est la plus proche. La figure 4 montre les vecteurs projetés (seules les expressions "Colère", "Tristesse" et "Dégoût" sont ici représentées) dans deux bases différentes : celle obtenue par une ACP classique suivie d'une ADL (image a), et celle obtenue après une ACP dont on a sélectionné les meilleures composantes suivie d'une ADL (image b). Nous remarquons que les trois classes sont plus compactes et mieux séparées si l'ACP a été suivie d'une sélection des meilleures composantes pour former une base de projection optimale.

4 Résultats et tests comparatifs

Nous avons testé notre méthode sur des images issues de la base du CMU-Pittsburgh. On construit l'ensemble

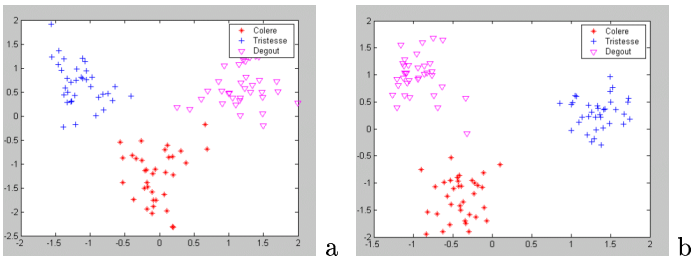

FIG. 4 – Séparabilité de trois classes après projection dans le sous-espace obtenu par : (a) une ACP classique suivie d'une ADL, (b) une ACP "optimisée" suivie d'une ADL.

TAB. 1 – Comparaison des erreurs classification après projection dans différentes bases (base CMU-Pittsburg).

Exp.	Eton.	Col.	Dég.	Joie	Peur	Trist.
#	81	41	80	82	54	35
B_1	2%	19%	5%	11%	27%	13%
B_2	4%	21%	5%	13%	27%	27%
B_3	2%	14%	5%	10%	9%	15%

d'apprentissage en utilisant les masques obtenus lors de l'extraction manuelle des traits caractéristiques (voir section 2). Le classifieur 6-expressions a été créé en apprenant les expressions sur un ensemble d'apprentissage composé de $N = 210$ masques faciaux (35 masques par expression) de personnes différentes (sexe, race). Ce classifieur a été testé avec des masques faciaux extraits d'images nouvelles (non apprises).

4.1 Pouvoir discriminant des sous-espaces de projection

Nous avons comparé trois différentes bases de projection : la première base (B_1) est construite avec les M premiers vecteurs propres (M étant le rang à partir duquel le taux d'inertie est supérieur à 0.9), la seconde base (B_2) est formée des K premiers (K étant le nombre optimal de composantes principales obtenu lors de l'étape de sélection, $K < M$) et la troisième base (B_3) des K correspondant aux "meilleures" composantes (classées par notre procédure). Le tableau 1 donne la comparaison des erreurs de classification, pour de nouvelles images après projection dans ces trois différentes bases : la supériorité des résultats obtenus avec la troisième base prouve l'intérêt de cette optimisation de l'espace de projection. La projection des individus dans la base "ACP optimisé+ADL" (selon les deux premières axes factoriels) nous a permis de constater que 3 expressions se distinguent des autres (étonnement, peur et joie), alors que les 3 autres sont assez proches, voire se recouvrent (colère, tristesse et dégoût).

4.2 Caractéristiques discriminantes pour la reconnaissance d'expressions

Nous avons comparé différentes parties du visage où l'expression s'exprime afin de connaître celles qui jouent le plus grand rôle dans la reconnaissance. Ainsi, nous avons étudié la zone des yeux (imagettes 1800 pixels), de la

TAB. 2 – Erreur moyenne de classification selon la zone caractéristique du visage considérée.

	Erreur moyenne
Yeux	42%
Bouche	18.2%
Yeux + Bouche	14.3%
Masque facial	11.7%

bouche (imagettes de 600 pixels), les imagettes des yeux et de la bouche (taille 2400) associées et enfin le masque facial étudié précédemment : des exemples de ces échantillons sont donnés par la figure 5. Pour toutes ces parties de visage, un classifieur 6-expressions a été entraîné par la procédure décrite dans la section 3. Les résultats de cette étude comparative ont été reportés dans le tableau 2 : l'erreur moyenne de classification (toutes expressions confondues) montre que les parties les moins discriminantes sont les yeux, et que le masque facial donne les meilleurs résultats. Dans le cadre de nos expérimentations, la partie interne est donc à prendre en compte dans son ensemble pour reconnaître l'expression faciale, et non seulement des parties isolées d'un visage.

FIG. 5 – Les deux traits caractéristiques isolés (de haut en bas) : les yeux (1800 pixels) et la bouche (600 pixels).

4.3 Représentation des masques faciaux

Le choix d'un type de représentation des masques joue un rôle important dans la qualité de discrimination. Nous avons testé 3 types de représentation : les niveaux de gris, le module du gradient de l'image et un module de Gabor de l'image. Ces représentations sont illustrées, pour chacune des 6 expressions universelles, par la figure 6. Nous donnons dans le tableau 3 les pourcentages d'erreur de classification en utilisant un classifieur 6-expressions pour les trois types de représentation. Les résultats montrent que l'utilisation des niveaux de gris minimise les erreurs de classification, même si, pour certaines expressions ("Dégoût" et "Joie"), l'information des frontières met plus en valeur l'expression. Les plus grandes erreurs sont obtenues avec le module de Gabor : cela vient du fait que nous n'avons choisi qu'une seule orientation ($\frac{\pi}{2}$) et résolution (π), alors qu'une seule ondelette ne permet pas de réagir de manière optimale avec toutes les expressions que l'on trouve dans l'ensemble d'apprentissage (mais nous voulions comparer des vecteurs de tailles identiques, ce qui n'aurait pas été le cas si nous avions effectué un filtrage multi-orientations et multi-résolutions). C'est pourquoi les méthodes de reconnaissance d'expressions faciales utilisant les fonctions de Gabor utilisent plusieurs ondelettes ([5], [7]), mais cela

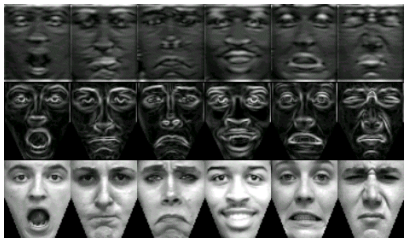

FIG. 6 – Trois représentations d'un même visage (de haut en bas) : module de Gabor, module du Gradient et les niveaux de gris pour les 6 expressions universelles, de gauche à droite, étonnement, colère, tristesse, joie, peur et dégoût.

TAB. 3 – Erreurs de classification pour 3 modes de représentation de visages (niveaux de gris, module de Gradient et module de Gabor).

Exp. #	Eton. 81	Col. 41	Dég. 80	Joie 82	Peur 54	Trist. 35
Niv. gris	2%	14%	5%	13%	9%	27%
Grad.	2%	24%	15%	17%	21%	7%
Gab.	19%	19%	22%	6%	18%	22%

augmente la taille des vecteurs d'entrée.

5 Conclusion

Nous avons présenté une méthode de reconnaissance d'expressions de visages optimisée par deux propriétés. La première consiste en une normalisation géométrique, et en luminance, de masques faciaux ne comprenant que la partie interne des visages. Ensuite, une méthode de choix optimiste de la base de projection des vecteurs d'apprentissage et de test permet d'améliorer à la fois la caractérisation et la séparabilité des classes. Nous avons montré l'amélioration des résultats, en terme de bonne classification, obtenue par l'utilisation de notre méthode. Enfin, des tests comparatifs ont permis d'établir que le masque facial en niveaux de gris semble être une représentation adaptée au problème de reconnaissance d'expressions faciales.

Références

[1] J. N. Bassili. Emotion recognition : The role of facial movement and the relative importance of upper and lower areas of the face. *J. Personality and Social Psychology*, 37 :2049–2059, 1979.

[2] G. Donato, M. Stewart Barlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE : Transactions on Pattern Analysis and Machine Intelligence*, 21(10) :974–989, oct 1999.

[3] P. Ekman and W.Friesen. *Facial Action Coding System : A Technique for the Measurement of Facial Movement*. Calif. : Consulting Psychologists Press, 1978.

[4] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, 1995.

[5] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE : Transactions on Pattern Analysis and Machine Intelligence*, 21(12) :1357–1362, dec 1999.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1) :71–86, 1991.

[7] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Computer Vision and Image Understanding*, 72(37) :286–296, 1998.

[8] Y. Yacoob and L.S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE : Transactions on Pattern Analysis and Machine Intelligence*, 18(6) :636–642, jun 1996.