

Modèles probabilistes d'apparence : une représentation approchée de faible complexité

A low complexity approximation of probabilistic appearance models

par R. HAMDAN, F. HEITZ ¹, L. THORAVAL

Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, UPRESA CNRS 7005/Université Strasbourg I, 4, bd Sébastien Brant, 67400 Illkirch, France.

¹ Auteur à contacter : Fabrice HEITZ, LSIT/CNRS, 4 bd Sébastien Brant, 67400 Illkirch, Tél. ++33 3 90 24 44 87, Fax ++33 3 90 24 43 42
e-mail : heitz@lsiit.u-strasbg.fr

résumé et mots clés

Les modèles d'apparence permettent d'encoder les variabilités de forme, de pose, et d'illumination dans une seule représentation compacte. Le modèle d'apparence probabiliste de Moghaddam et al. (Moghaddam and Pentland, 1997; Tipping and Bishop, 1997b), reposant sur une interprétation statistique de l'Analyse en Composantes Principales (ACP) s'est récemment illustré par ses excellentes performances en détection et en reconnaissance des formes, surpassant de nombreuses autres méthodes linéaires et non linéaires. Ce modèle, performant, se heurte toutefois à une complexité calculatoire importante. Nous proposons, dans cet article, une approximation de ce modèle qui se prête à une mise en œuvre rapide, dans le cadre de schémas d'estimation statistique. Des gains en complexité et en temps de calcul supérieurs à 10, sont obtenus, sans aucune perte de qualité dans les résultats des traitements.

Modèles d'apparence, modèles probabilistes, algorithmes rapides, détection, maximum de vraisemblance.

abstract and key words

Appearance models yield a compact representation of shape, pose and illumination variations. The probabilistic appearance model, proposed by Moghaddam et al. (Moghaddam and Pentland, 1997; Tipping and Bishop, 1997b), has recently shown excellent performances in pattern detection and recognition, outperforming most other linear and non-linear approaches. Unfortunately, the complexity of this model remains high. In this paper, we introduce an efficient approximation of this model, which enables fast implementations in statistical estimation-based schemes. Gains in complexity and cpu time of more than 10 have been obtained, without any loss in the quality of the results.

Appearance models, probabilistic models, fast algorithms, detection, maximum likelihood

1. introduction

Les premiers modèles d'objets en vision par ordinateur ont mis en relief les descripteurs géométriques 2D/3D de la forme (Lowe, 1991), tels que les contours ou les éléments de surface. De tels descripteurs permettent une représentation compacte, faiblement dense de l'information. Ils présentent une faible sensibilité aux changements d'illumination et sont peu coûteux à extraire. Leur extraction (segmentation) et leur mise en correspondance avec les indices de l'image 2D/3D fait toutefois l'objet de recherches encore actives, même après plus de trente années de travaux sur le sujet. Ce type de descripteur est bien adapté pour la description d'objets manufacturés dans des environnements « artificiels », riches en contours. Dans des environnements moins structurés (milieux naturels, objets biologiques, images médicales, etc.), les contours sont présents de façon plus éparse et moins prédictible : ils sont plus difficiles à détecter et à mettre en correspondance avec des modèles.

C'est pourquoi, on a assisté, au début des années 1990, à un regain d'intérêt pour les modèles d'apparence (Murase et Nayar, 1995; Turk et Pentland, 1991), auxquels se rattachent des approches anciennes de la reconnaissance des formes, comme la mise en correspondance de gabarits (« template matching ») ou le filtrage adapté (Jain *et al.*, 2000). Un modèle d'apparence permet en effet d'encoder à la fois la forme, la pose, et les variations d'illumination dans une seule représentation compacte, en particulier grâce aux techniques de réduction de dimension de l'analyse des données (Saporta, 1992) ou de la reconnaissance statistique des formes (Fukunaga, 1990). Les modèles d'apparence représentent les objets à partir de leurs apparences dans le domaine 2D de l'image en offrant la possibilité d'exploiter directement l'information brute : l'intensité des pixels (Murase et Nayar, 1995; Turk et Pentland, 1991), ou des informations de bas niveau directement dérivées de l'intensité (dérivées spatiales, « jets » (Schmid et Mohr, 1997)). Ils définissent ainsi une approche alternative à la description éparse par primitives géométriques, particulièrement utile lorsque l'environnement ou les objets à modéliser sont peu structurés.

Les premiers succès des modèles d'apparence globaux, en particulier en reconnaissance des visages (Turk et Pentland, 1991) ont suscité de nombreux travaux qui ont, entre autres, conduit à la possibilité de reconnaître des objets 3D dans des bases de plus de 100 objets (Murase et Nayar, 1995) (dans des environnements contrôlés), puis plus récemment la reconnaissance, avec un réel succès, de classes très générales d'objets (voitures, visages) (Schneiderman et Kanade, 2000) dans des environnements complexes.

Dans cet article nous nous intéressons à une classe particulière de modèles d'apparence : les modèles probabilistes d'apparence globaux introduits récemment, conjointement et indépendamment par Moghaddam et Pentland (Moghaddam et Pentland,

1997) et Tipping et Bishop (Tipping et Bishop, 1997c). Les modèles de Moghaddam *et al.* présentent quelques avantages déterminants dans les applications :

- ils sont probabilistes : ils permettent de représenter une classe d'images et rendent accessibles toutes les méthodes classiques de l'estimation statistique (maximum de vraisemblance, approches bayésiennes). Comme les réseaux neuromimétiques, ils possèdent des capacités de « généralisation », en affectant une densité de probabilité à toute image de l'espace représenté ;
- ce sont des modèles paramétriques (spécifiés par un faible nombre de paramètres, contrairement aux histogrammes) ;
- ces modèles sont linéaires (globalement ou par morceaux), et se prêtent donc à des implantations efficaces ;
- bien qu'étant linéaires, ils se sont révélés supérieurs, en termes de détection et reconnaissance, non seulement aux méthodes linéaires classiques (Analyse en Composantes Principales – ACP –, Analyse en Composantes Indépendantes, etc.) mais également aux approches non linéaires (réseaux neuromimétiques, ACP non linéaire), dans une comparaison récente menée par Moghaddam et Pentland (Moghaddam, 1999). Une méthode bayésienne, utilisant ces modèles a également surpassé toutes les approches connues de reconnaissances de visages sur la base de données « FERET » (Moghaddam *et al.*, 1999) ;
- ils se prêtent enfin à des extensions non linéaires basées sur des modèles statistiques robustes, qui permettent d'en améliorer encore les performances (Dahyot *et al.*, 2001).

Le modèle global d'apparence considéré s'appuie sur une interprétation probabiliste des techniques d'analyse en composantes principales (Tipping et Bishop, 1997b ; Tipping et Bishop, 1997a), en représentant la distribution du nuage de points, correspondant aux images d'une base d'apprentissage, par un modèle statistique paramétrique (gaussien ou multi-gaussien).

Ce modèle, performant, se heurte toutefois à une complexité calculatoire importante, qui interdit, pour l'instant, son utilisation dans des applications exigeantes en temps de calcul (extraction et suivi de formes dans une séquence d'images, indexation de bases d'images de grande dimension, etc.). Nous proposons, dans cet article, une approximation du modèle d'apparence gaussien original de Moghaddam *et al.* qui autorise une mise en œuvre rapide de ce modèle, dans le cadre de schémas d'estimation statistique. L'approximation repose sur une Décomposition en Valeurs Singulières (SVD) des vecteurs propres et du vecteur moyen associés au modèle gaussien. La vraisemblance du modèle résultant peut s'évaluer de façon rapide, sous forme de corrélations séparables (mono-dimensionnelles). Des gains en complexité et en temps de calcul significatifs, de l'ordre de 10, sont obtenus grâce à cette approche, sans aucune perte de qualité dans les résultats des traitements. Cette approche se généralise par ailleurs de façon immédiate au cas des modèles multi-gaussiens (Moghaddam et Pentland, 1997).

La suite de l'article est organisée comme suit. Dans la partie 2, nous présentons tout d'abord le modèle gaussien original de Moghaddam *et al.* (Moghaddam et Pentland, 1997) et Tipping *et al.* (Tipping et Bishop, 1997b) ainsi que les techniques mises en œuvre pour l'estimation de ses paramètres (apprentissage). Nous introduisons ensuite, dans la partie 3, l'approximation du modèle original, reposant sur la décomposition en valeurs singulières (SVD) des vecteurs propres et du vecteur moyen du modèle. Cette approximation se prête à une implantation rapide, dont la complexité est comparée au modèle initial. Une application à la détection, par maximum de vraisemblance, de formes d'intérêt (dont l'apparence est modélisée), permet de comparer expérimentalement les performances du modèle approché, par rapport au modèle original (partie 4). Des gains en temps de calcul de l'ordre 10 peuvent être obtenus, sans modification des performances de détection.

2. modèle probabiliste d'apparence

Moghaddam et Pentland (Moghaddam et Pentland, 1997) ainsi que Tipping et Bishop (Tipping et Bishop, 1997a; Tipping et Bishop, 1997b) ont développé des modèles probabilistes de l'apparence, sous la forme de distributions gaussiennes ou multi-gaussiennes, intégrant les techniques linéaires de réduction de dimension (ACP ou transformée de Karhunen-Loeve). Le modèle de distribution gaussienne de Moghaddam et Pentland, contrairement aux approches simples par ACP, modélise non seulement la distribution des exemples d'apprentissage sur le sous-espace propre de dimension réduite associé à l'ACP, mais approche également la distribution de cet ensemble sur l'espace orthogonal (complémentaire), ce qui fournit une description complète. Ceci permet de discriminer des images dont les composantes ne diffèrent que sur l'espace orthogonal et d'utiliser les approches classiques en détection et estimation statistique (maximum de vraisemblance, approches bayésiennes, etc.).

2.1. définition du modèle et apprentissage

Dans une phase préalable d'apprentissage, on réunit un ensemble représentatif d'images en niveaux de gris \mathbf{x} , de dimension $N = L^2$ pixels (rangés par exemple dans l'ordre lexicographique), présentant les différentes apparences 2D de la structure à modéliser (voir par exemple Fig.1a - on considère ici, pour simplifier les notations, des images « carrées »). Pour assurer une certaine robustesse de la représentation aux variations d'illumination globale ainsi qu'aux conditions d'acquisition, chaque image de l'ensemble d'apprentissage (et chaque imagerie analysée) est normalisée en moyenne et variance pour obtenir un

signal aléatoire centré (de moyenne nulle) et de variance unité. La même normalisation sera appliquée aux observations, lorsque le modèle sera utilisé.

Les statistiques du premier et du second ordre (moyenne μ et matrice d'autocovariance \mathbf{Q}) sont ensuite estimées, par des moyennes d'ensemble classiques (valeurs moyennes statistiques), à partir de cet ensemble d'apprentissage. L'image moyenne μ est simplement obtenue en moyennant l'ensemble des N_T images \mathbf{x}_i de la base d'apprentissage :

$$\mu = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{x}_i$$

Pour estimer la matrice de covariance, on vient former la matrice \mathbf{X} , de dimensions $N \times N_T$:

$$\mathbf{X} = [\mathbf{x}_1 - \mu, \mathbf{x}_2 - \mu, \dots, \mathbf{x}_{N_T} - \mu]$$

L'estimée de la matrice d'autocovariance (de taille $N \times N$) s'écrit alors classiquement :

$$\mathbf{Q} = \frac{1}{N_T} \mathbf{X} \mathbf{X}^T$$

Le modèle de Moghaddam *et al.*, qui ne prend en compte que les statistiques jusqu'au second ordre (moyenne, autocovariance) sous-tend une distribution gaussienne des données, pour élaborer un modèle d'apparence à N variables $\mathcal{N}(\mathbf{x}|\mu, \mathbf{Q})$. La vraisemblance d'une image \mathbf{x} (de dimension N) s'exprime donc par :

$$P(\mathbf{x}|\mu, \mathbf{Q}) = \frac{\exp[-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{Q}^{-1}(\mathbf{x} - \mu)]}{(2\pi)^{N/2} |\mathbf{Q}|^{1/2}} \quad (1)$$

L'utilisation directe d'un tel modèle gaussien pose plusieurs problèmes :

- Le nombre N_T d'exemples d'apprentissage est généralement très inférieur à la dimension de l'espace des images. Il en résulte une matrice d'autocovariance singulière (de rang incomplet), ce qui pose évidemment problème pour son inversion !
- Même si la matrice \mathbf{Q} était de rang complet, une évaluation directe de la vraisemblance (1) est très coûteuse, en raison de la dimension du vecteur image \mathbf{x} (N est de l'ordre de 100×100 dans notre cas).

En pratique, les images du corpus d'apprentissage sont très corrélées. Une décorrélation de ces images, par ACP ou transformée de Karhunen-Loeve (TKL) du vecteur aléatoire \mathbf{x} , permet classiquement de réduire la dimension du problème (Fukunaga, 1990) et de se concentrer (en supprimant les valeurs propres nulles) sur le sous-espace où la matrice \mathbf{Q} est de rang non déficient. Le calcul de la TKL implique la diagonalisation de la matrice de covariance :

$$\mathbf{Q} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$$

où \mathbf{U} désigne la matrice des vecteurs propres, et Λ est la matrice diagonale des valeurs propres. La TKL est alors définie par la projection :

$$\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \mu) \quad (2)$$

Avec ce changement de variable, la fonction de la vraisemblance (1) peut s'écrire :

$$P(\mathbf{x}|\mu, Q) = \frac{\exp\left[-\frac{1}{2} \sum_{i=1}^N \frac{y_i^2}{\lambda_i}\right]}{(2\pi)^{N/2} |\Lambda|^{1/2}} \quad (3)$$

où les y_i désignent les composantes du vecteur \mathbf{y} et les λ_i sont les valeurs propres (les éléments diagonaux de la matrice Λ). Remarquons que l'expression précédente devient singulière si les valeurs propres λ_i s'annulent au-delà d'un certain rang (ce qui est presque toujours le cas, comme nous l'avons indiqué : c'est cette propriété qui permet de réduire la dimension du problème).

La réduction de la dimension du problème est obtenue en approchant $P(\mathbf{x}|\mu, Q)$ à partir des M ($M \ll N$) composantes principales associées à la TKL. En pratique on ne laisse pas uniquement de côté les valeurs propres nulles, mais également celles de faible valeur.

Moghaddam et Pentland (Moghaddam et Pentland, 1997) proposent, sur ce principe, une approximation « optimale » de $P(\mathbf{x}|\mu, Q)$, sur le sous-espace engendré par les M premiers vecteurs propres et *sur son orthogonal* :

$$\hat{P}(\mathbf{x}|\mu, Q) = \left[\frac{\exp\left(-\sum_{i=1}^M \frac{y_i^2}{2\lambda_i}\right)}{(2\pi)^{M/2} \prod_{i=1}^M \lambda_i^{1/2}} \right] \left[\frac{\exp\left(-\sum_{i=M+1}^N \frac{y_i^2}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}} \right] \quad (4)$$

Le premier terme correspond au terme gaussien classique associé à une ACP retenant les M composantes avec les valeurs propres les plus fortes. Le second terme constitue une *approximation non singulière* de la distribution sur l'espace orthogonal (engendrée par les $N - M$ autres composantes, correspondant aux valeurs propres les plus petites). Cette approximation est optimale dans le sens où le paramètre ρ est estimé pour minimiser la distance de Kullback-Leibler entre $P(\mathbf{x}|\mu, Q)$ et $\hat{P}(\mathbf{x}|\mu, Q)$ (Moghaddam et Pentland, 1997) :

$$\rho = \frac{1}{N - M} \sum_{i=M+1}^N \lambda_i$$

Tipping et Bishop (Tipping et Bishop, 1997b) obtiennent le même modèle, en adoptant une interprétation probabiliste de l'ACP, associée à une estimation au sens du maximum de vraisemblance des paramètres du modèle.

L'intérêt de l'expression obtenue pour la distribution sur l'espace orthogonal est que le résiduel de reconstruction dans cet espace

orthogonal, peut être calculé très efficacement à partir des $M \ll N$ premières composantes principales :

$$\sum_{i=M+1}^N y_i^2 = \|\mathbf{x} - \mu\|^2 - \sum_{i=1}^M y_i^2 \quad (5)$$

Un exemple de vecteur moyen et les premiers vecteurs propres associés au modèle probabiliste sont présentés, à titre d'illustration, figure 1, pour un échantillon d'apprentissage correspondant à des images de mains, dans différentes configurations, correspondant à l'exécution de gestes simples.

2.2. remarques sur l'apprentissage

En pratique, la matrice d'autocovariance est construite à partir d'un nombre limité N_T d'observations. Le nombre de valeurs propres non nulles est au plus égal au nombre d'exemples d'apprentissage moins 1. Les valeurs propres au-delà d'un certain rang étant faibles et entachées par le bruit, Moghaddam & Pentland (Moghaddam et Pentland, 1997) proposent d'ajuster une fonction de la forme $(1/x)$ sur les valeurs propres estimées entre M et N . C'est la solution que nous avons retenue pour le calcul de ρ .

Par ailleurs, pour calculer les vecteurs propres de la matrice de covariance \mathbf{Q} , nous n'effectuons pas une diagonalisation directe de $\mathbf{Q} = \frac{1}{N_T} \mathbf{X} \mathbf{X}^T$ (qui serait très coûteuse, car cette matrice est de très grande taille $N \times N$), mais nous nous appuyons sur une méthode indirecte, reposant sur la diagonalisation de $\mathbf{X}^T \mathbf{X}$ de dimensions $N_T \times N_T$ (voir (Golub et Loan, 1989; Murase et Lindenbaum, 1995)).

3. modèle probabiliste approché

3.1. décomposition en valeurs singulières

La décomposition en valeurs singulières (SVD) d'une matrice, est une technique très utile en traitement du signal. Une synthèse de l'utilisation de la SVD en traitement des signaux et en identification de systèmes est proposée dans (Van-Der-Veen *et al.*, 1993).

Dans notre problème, une approximation du modèle de Moghaddam et Pentland a été élaborée en utilisant la décomposition en valeurs singulières de la forme matricielle de la moyenne et des vecteurs propres du modèle d'apparence (4). Une projection (corrélation) séparable avec les premières com-

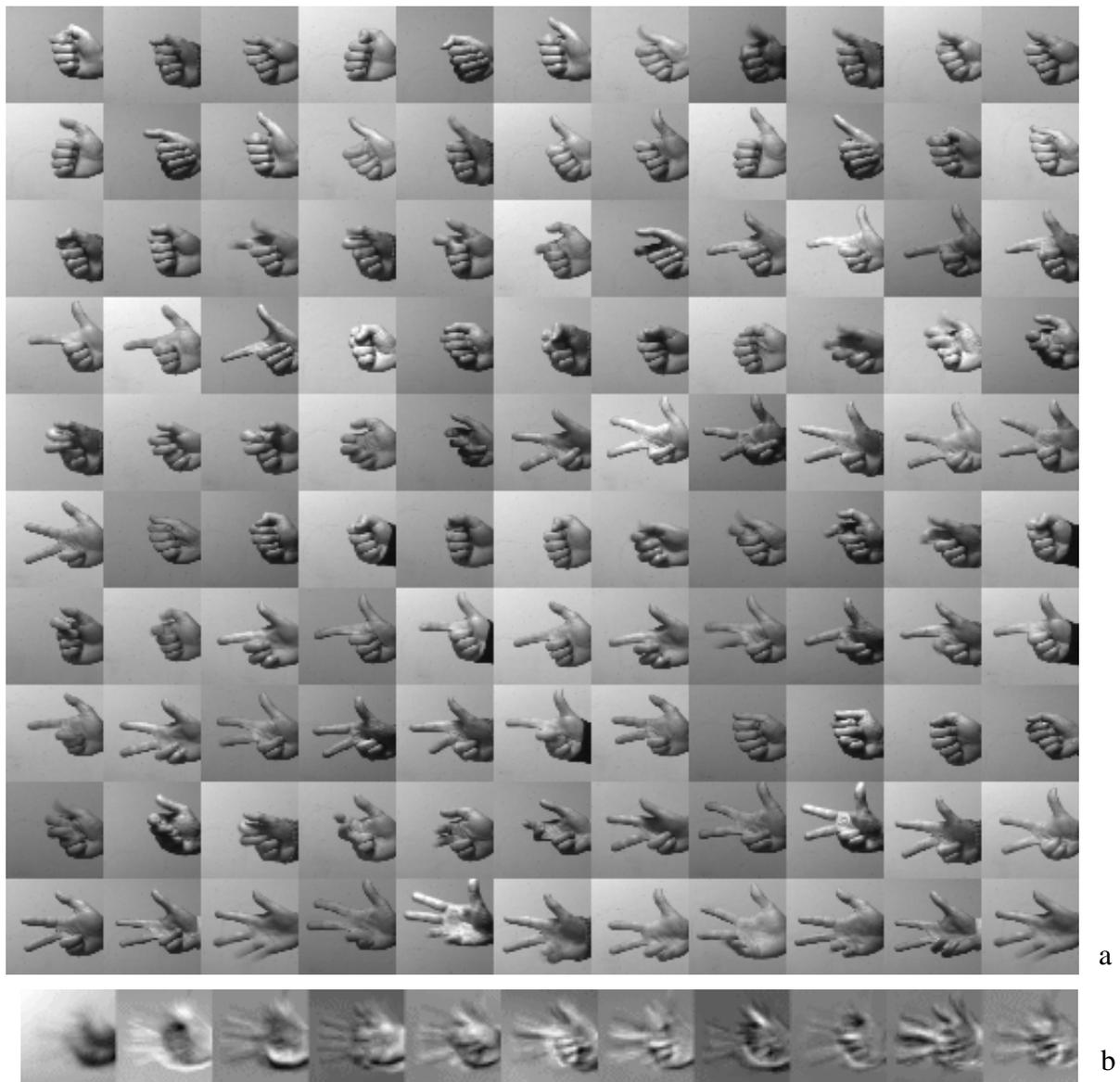


Figure 1. - (a) Exemples de l'ensemble d'apprentissage. (b) Moyenne et premiers vecteurs propres associés au modèle d'apparence de la main.

posantes de la SVD de chacun de ces vecteurs permet d'accélérer de façon significative les calculs, sans perte importante d'information.

Rappelons qu'en traitement du signal, une fonction $f(x, y)$ est dite séparable s'il est possible de la factoriser sous la forme $f(x, y) = f_1(x) f_2(y)$. Le caractère séparable de certaines fonctions à deux variables permet de simplifier grandement le traitement des signaux à deux dimensions. C'est en particulier le cas des filtres dits « séparables », dont la réponse impulsionnelle est supposée vérifier la propriété de séparabilité. De tels filtres peuvent être implantés (par convolution dans le domaine spatial) en effectuant successivement un filtrage monodimensionnel selon les lignes et les colonnes de l'image, opération beaucoup

moins coûteuse qu'un filtrage par un noyau de convolution 2D. Un filtre bidimensionnel quelconque n'est en général pas séparable, mais la décomposition en valeurs singulières permet de l'approcher par une somme de filtres séparables. Nous nous sommes inspirés de cette propriété pour développer notre approximation (qui n'est toutefois pas triviale, en particulier en raison de la normalisation des observations).

La décomposition en valeurs singulières d'une matrice \mathbf{A} de dimension $m \times n$ ($m \leq n$) consiste à la factoriser sous la forme :

$$\mathbf{A} = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (6)$$

où : \mathbf{U} est une matrice unitaire de taille $m \times m$ dont les colonnes sont les vecteurs propres de $\mathbf{A}\mathbf{A}^T$; \mathbf{V} est la matrice

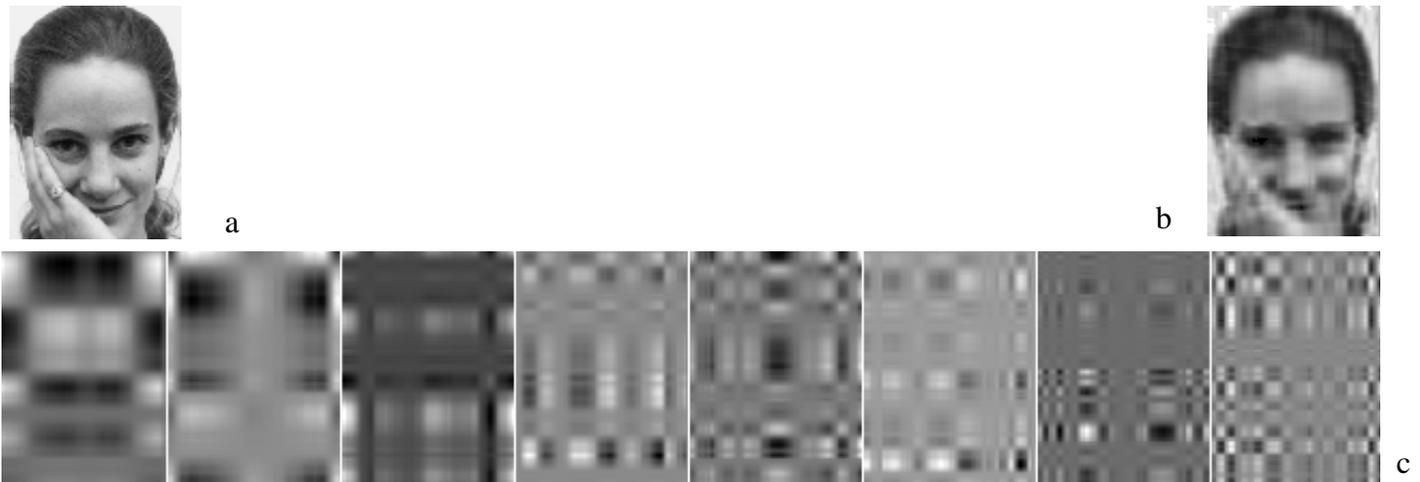


Figure 2. – (a) Image originale. (b) La somme des 8 premières composantes de la SVD. (c) Les 8 premières composantes ou « images propres ».

unitaire $n \times n$ dont les colonnes sont les vecteurs propres de $\mathbf{A}^T \mathbf{A}$; \mathbf{S} est la matrice diagonale $m \times n$ des valeurs singulières ψ_i qui sont les racines carrées des valeurs propres de $\mathbf{A}^T \mathbf{A}$ ou $\mathbf{A} \mathbf{A}^T$. On a $r < m$ valeurs singulières non nulles, si r est le rang de la matrice \mathbf{A} .

Dans le cas où \mathbf{A} représente la réponse impulsionnelle d'un filtre, cette décomposition permet un filtrage séparable. Étant donné que la matrice \mathbf{S} est diagonale, l'équation (6) peut être développée en une somme de matrices de rang 1 ($\mathbf{u}_i \mathbf{v}_i^T$) pondérées par les ψ_i :

$$\mathbf{A} = \sum_{i=1}^r \psi_i \mathbf{u}_i \mathbf{v}_i^T$$

où \mathbf{u}_i et \mathbf{v}_i sont les vecteurs colonnes des matrices \mathbf{U} et \mathbf{V} , correspondant à la valeur singulière ψ_i . En classant les valeurs singulières dans l'ordre décroissant, une bonne approximation de la matrice \mathbf{A} , au sens de la norme de Frobenius (Van-Der-Veen *et al.*, 1993), peut être élaborée en examinant le spectre des valeurs singulières. Une somme tronquée aux $k \ll r$ premières composantes, donne l'approximation optimale \mathbf{A}_k de la matrice originale par une somme de matrices de rang 1 :

$$\mathbf{A}_k = \sum_{i=1}^k \psi_i \mathbf{u}_i \mathbf{v}_i^T$$

Le carré de la norme de l'erreur d'approximation entre \mathbf{A} et \mathbf{A}_k est égal à (Andrews et Hunt, 1977; Van-Der-Veen *et al.*, 1993) :

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i=k+1}^r \psi_i^2 \quad (7)$$

où $\|\cdot\|_F$ est la norme de Frobenius de la matrice, définie par : $\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^T \mathbf{A})$ (il s'agit simplement de la somme des carrés de ses éléments).

En général, des valeurs de k de l'ordre de quelques unités permettent déjà une excellente approximation pour des matrices représentant des images numériques (Andrews et Hunt, 1977).

A titre d'illustration, la figure 2 présente la décomposition en SVD d'une image de visage. La somme des huit premières composantes de la SVD, fournit déjà une bonne approximation de l'image de départ (l'image étant de taille 80×110 , on a dans ce cas 80 valeurs singulières). Le spectre des valeurs singulières (c'est-à-dire la représentation des ψ_i en fonction de i), figure 3, montre cette concentration de l'information dans quelques valeurs singulières dominantes. Nous avons observé la même propriété pour la moyenne et les vecteurs propres qui interviennent dans notre modèle gaussien, pour lesquels un faible nombre

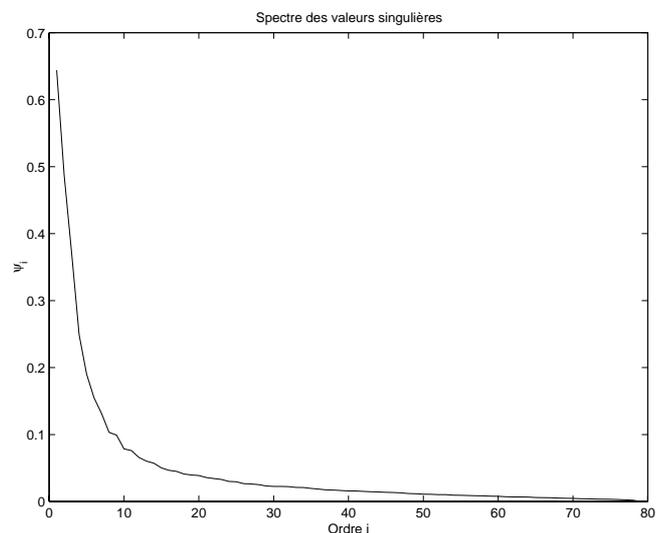


Figure 3. – Spectre des valeurs singulières de l'image de la figure 2.

de composantes permet de préserver l'essentiel de l'information et d'obtenir d'excellentes performances en détection ou reconnaissance des formes. L'image moyenne du modèle, ainsi que les premiers vecteurs propres de l'ACP sont en effet des images « lisses », puisque représentatives de nombreuses variabilités. Ces images se prêtent donc à une compression importante d'information par SVD (entre 2 et 10 valeurs singulières permettent de retenir plus de 98 % de l'information initiale, comme nous le verrons dans la suite).

3.2. convolution séparable

La convolution séparable d'un filtre, représenté par sa réponse impulsionnelle \mathbf{A} , avec une image, est effectuée de la façon suivante. La SVD est tout d'abord appliquée à \mathbf{A} , en retenant les composantes correspondant aux $k \ll r$ valeurs singulières ψ_i les plus élevées (en pratique les valeurs singulières ψ_i sont intégrées dans les vecteurs $\mathbf{u}_i, \mathbf{v}_i$ ce qui donne des vecteurs pondérés). Une convolution monodimensionnelle des colonnes de l'image est réalisée avec la première colonne de \mathbf{U} (vecteur \mathbf{u}_1), puis les lignes de l'image résultante sont convoluées avec le vecteur \mathbf{v}_1^T . La même procédure est appliquée pour la deuxième composante de la SVD $\mathbf{u}_2 \mathbf{v}_2^T$ et ainsi de suite jusqu'à l'ordre k . La somme des résultats intermédiaires constitue alors une bonne approximation de la convolution avec le filtre \mathbf{A} .

Une convolution directe avec un filtre discret de $L \times L$ points nécessite L^2 opérations (multiplications + additions) pour chaque pixel de l'image. Pour la convolution séparable, seules $2L$ opérations sont nécessaires pour chaque pixel et chaque composante de la SVD. La convolution avec les premières $k \ll r < L$ composantes nécessite donc $2kL \ll L^2$ opérations. Dans notre cas, l'évaluation du premier terme du modèle (4) implique M corrélations avec les M premiers vecteurs propres du sous-espace propre principal E , chaque vecteur propre étant de taille $L^2 = N$. Ces corrélations, qui doivent être calculées pour chaque point de l'image dans laquelle la structure modélisée est recherchée, peuvent se réécrire comme des convolutions avec des masques de taille $L \times L$. Notre approche consiste donc à mettre les M vecteurs propres sous forme matricielle, puis à les factoriser en SVD pour exploiter l'avantage de la convolution séparable. Cette factorisation se fait hors-ligne puisque les vecteurs propres sont déterminés, préalablement dans la procédure d'apprentissage.

La corrélation avec chaque vecteur propre du sous-espace propre E nécessite $O(L^2)$ opérations par pixel, alors que la méthode séparable ne requiert que $O(2kL)$ opérations, si $k \ll L$ composantes sont retenues dans la SVD. L'économie en temps de calcul est donc substantielle. Par ailleurs, l'approche est bien adaptée pour une implantation parallèle (machine multiprocesseur) car la corrélation avec les différentes composantes peut être menée en parallèle, avec un accumulateur pour les résultats.

Un gain supplémentaire en complexité calculatoire pourrait en principe être obtenu en utilisant la transformée de Fourier rapide (FFT) pour implanter les corrélations (ou convolutions) 1D, lorsque la taille des vecteurs 1D devient importante. Ceci n'est toutefois pas possible en pratique, en raison de la normalisation glissante des observations extraites de l'image (en moyenne et variance), nécessaire pour obtenir une certaine robustesse aux conditions d'illumination. Cette normalisation glissante conduit à un système non linéaire et non stationnaire (voir Annexe A).

3.3. complexité comparée des modèles

La complexité du calcul de la vraisemblance pour le modèle original de Moghaddam *et al.* et pour notre modèle approché est évaluée de façon détaillée dans l'annexe A.

Considérons un modèle d'imagette de hauteur et de largeur L , ce qui correspond à un vecteur d'observation de dimension $N = L^2$ pixels. Cette fenêtre d'observation est déplacée dans une zone de recherche de taille $P \times P$ (qui est définie dans l'image dans laquelle on cherche la forme modélisée, voir Fig. 4). On est ainsi amené à évaluer P^2 fois la vraisemblance du modèle.

On évalue le nombre d'opérations nécessaires pour évaluer la vraisemblance, pour une position de l'imagette dans la zone de recherche.

Pour M vecteurs propres retenus dans l'ACP, l'évaluation de la log-vraisemblance du modèle original de Moghaddam *et al.* (4) nécessite $(M + 2)N$ multiplications et $(M + 5)N$ additions, pour chaque position p de la fenêtre d'analyse (voir Annexe A). L'implantation adoptée pour le modèle séparable (non triviale, car nécessitant de gérer la normalisation préalable des observations en moyenne et variance) permet de ramener le nombre d'opérations à $2k(M + 1)L$ multiplications et additions, où k

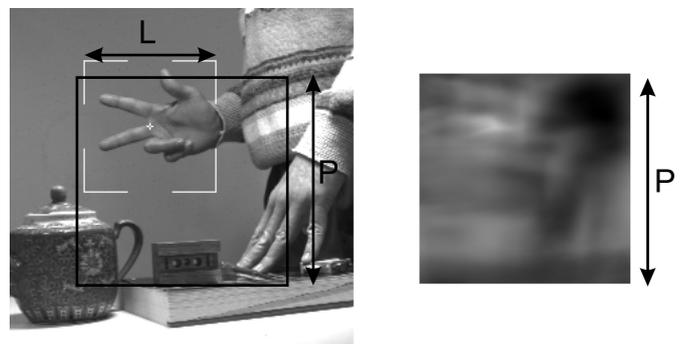


Figure 4. – Détection d'une forme modélisée et carte de vraisemblance associée (modèle statistique de main). A gauche : image originale, avec en superposition la fenêtre correspondant à la position estimée de la main. A droite : carte de la log-vraisemblance des observations (Equ. 4). Le maximum (point le plus clair) correspond à la position estimée de la main, centre de la fenêtre blanche affichée sur l'image originale. L^2 : dimension du modèle, P^2 : dimension de la zone de recherche.

Tableau 1. – Comparaison des complexités et temps de calcul pour l'implantation séparable et standard pour différents modèles et zones de recherche ($M = 5$ et $k = 6$ dans tous les cas).

taille modèle L	30	30	75	99
taille zone de recherche P	128	200	256	256
implantation standard	10 s	45 s	280 s	360 s
implantation séparable	4 s	15 s	30 s	35 s
rapport temps cpu	3	3	9	10
rapport complexité r_{mult}	3	3	7	10
rapport complexité r_{add}	4	4	10	14

désigne le nombre de valeurs singulières conservées dans l'approximation (pour les détails, le lecteur est renvoyé à l'annexe A).

Le rapport des complexités du modèle original et du modèle séparable est donc, pour le nombre de multiplications :

$$r_{mult} = \frac{(M + 2)L}{2(M + 1)k}$$

pour les additions :

$$r_{add} = \frac{(M + 5)L}{2(M + 1)k}$$

Le tableau 1 présente les rapports de complexité et temps de calcul obtenus sur une station de travail (200 Mhz), pour différentes tailles d'images et de modèle. La zone de recherche correspond dans ce cas à toute l'image, aux effets de bord près. On remarque que l'on observe un bon accord entre les complexités théoriques et le temps cpu mesuré.

Dans notre application, nous avons typiquement les paramètres $M = 5, L = 100, k = 5$. Le rapport des complexités est donc d'environ $r_{mult} \simeq 10$ et $r_{add} \simeq 15$, dans notre cas.

4. application à la détection d'objets

4.1. détection au sens du maximum de vraisemblance

Dans cette partie nous présentons une application du modèle à la détection de formes modélisées, au sens du maximum de vraisemblance (Moghaddam et Pendland, 1997). Cette approche est appliquée au modèle original de Moghaddam *et al.* (Moghaddam et Pendland, 1997), ainsi qu'à notre modèle approché, qui

se distingue par de bonnes performances en détection (comparables au modèle original) et, comme annoncé, un gain significatif en temps de calcul.

Une fois que l'on a construit un modèle d'apparence caractérisé par sa densité de probabilité $\hat{P}(\mathbf{x}|\mu, Q)$ (4), la détection de la forme est obtenue par une estimation au sens du maximum de vraisemblance, comme suit. Une fenêtre glissante de $N = L^2$ pixels (de même taille et proportion que les exemples d'apprentissage), centrée au point (i, j) balaye l'ensemble de l'image. Le vecteur d'observation est constitué par le vecteur $\mathbf{x}^{(i,j)}$ des pixels de la fenêtre, ordonnés selon l'ordre lexicographique. La segmentation est obtenue en détectant la position de la fenêtre (i, j) conduisant à la vraisemblance maximale du vecteur d'observation, selon le modèle (4) (ou sa version approchée « séparable ») :

$$\begin{aligned} (\widehat{i, j})_{opt} &= \arg \max_{(i,j)} \hat{P}(\mathbf{x}^{(i,j)}|\mu, Q) \\ &= \arg \max_{(i,j)} \left[\frac{\exp\left(-\sum_{l=1}^M \frac{y_l^{(i,j)^2}}{2\lambda_l}\right)}{(2\pi)^{M/2} \prod_{l=1}^M \lambda_l^{1/2}} \right] \left[\frac{\exp\left(-\sum_{l=M+1}^N \frac{y_l^{(i,j)^2}}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}} \right] \end{aligned} \quad (8)$$

où le vecteur $\mathbf{y}^{(i,j)}$ est obtenu à partir de l'observation $\mathbf{x}^{(i,j)}$ par la transformation de Karhunen-Loeve (2).

La figure 4 montre un exemple de détection-localisation et la carte de log-vraisemblance associée pour un modèle statistique de main. Cet exemple permet d'apprécier qualitativement la bonne localisation obtenue par cette approche (pour d'autres exemples, voir (Hamdan, 2001 ; Moghaddam et Pentland, 1997)).

4.2. choix du nombre de vecteurs propres

Un choix important, dans la construction du modèle, est celui du nombre M de vecteurs propres retenus pour représenter l'espace propre principal $E = \{\mathbf{U}\}_1^M$; les $N - M$ autres vecteurs étant affectés à son orthogonal \bar{E} . Le choix usuellement adopté pour des tâches de reconnaissance des formes (dans des situations où toutes les images observées sont *dans* le sous-espace propre principal E) consiste à modéliser de façon la plus complète possible E en retenant les M vecteurs propres principaux selon des critères simples de préservation d'information (Murase et Nayar, 1995). La quantité d'information conservée est directement mesurable à partir de la somme des valeurs propres associées (Fukunaga, 1990). On retient ainsi typiquement un nombre de vecteurs propres correspondant à 90 ou 95 % de l'information totale. Dans notre cas (détection - discrimination entre images d'objets et de « non objets »), le rôle du terme dans l'espace

orthogonal est important et une série d'expériences avec différentes « ventilations » des vecteurs propres entre E et \bar{E} nous ont conduit à affecter la moitié de l'information expliquée (soit 50 % de la somme des valeurs propres) à E , l'autre moitié à \bar{E} . Ce résultat a été obtenu par des essais de détection sur de grandes bases de données de visages et de mains, qui sont présentés dans la suite. Ce point mériterait toutefois une étude théorique plus approfondie (des techniques bayésiennes de sélection de l'ordre de modèles s'appuyant sur le modèle probabiliste d'ACP de Tipping et Bishop (Tipping and Bishop, 1997a) ont été développées récemment dans (Minka, 1999)).

4.3. résultats expérimentaux

L'implantation de l'estimateur du maximum de vraisemblance pour le modèle de Moghaddam *et al.* ainsi que pour notre version approchée rapide a été testée et validée sur une collection de plus de 5000 images, pour la détection de primitives du visage comme l'œil et la bouche, ainsi que pour la segmentation de structures articulées comme les mains.

4.3.1. bases de données

Les images de visages, que nous avons considérées, appartiennent à différentes bases de données disponibles sur le réseau :

- la base de visages de l'Université de Yale (<http://cvc.yale.edu/projects/yalefaces/yalefaces.html>), qui contient les images de 15 personnes, chaque personne étant représentée par 10 images prises avec différentes conditions d'illumination et différentes expressions faciales ;
- la base de visages du Laboratoire de Recherche d'Olivetti, Cambridge, UK., qui contient 400 images de 40 personnes. Les dix images représentant chaque personne ont été acquises dans des conditions d'illumination variables, avec diverses expressions faciales, avec ou sans lunettes ;
- la base de données de visages de l'Université de Manchester (Equipe du Pr. Taylor) (<http://peipa.essex.ac.uk/ftp/ipa/pix/faces/manchester>), dont nous avons extrait une centaine d'images.

Des ensembles d'apprentissage pour les structures « œil gauche » et « bouche » ont été constitués par segmentation semi-manuelle à partir de 50 imageries pour chaque modèle. Les figures 5 et 6 présentent quelques exemples d'imageries d'apprentissage pour l'œil, la bouche, ainsi que l'image moyenne et les vecteurs propres du modèle d'apparence associé.

Nous avons également constitué dans notre laboratoire, une base de données de 200 séquences vidéo (25 images/s) de 5 gestes de la main, réalisés par 8 personnes différentes. Les séquences varient en longueur entre 25 et 45 images et ont une résolution de 256×256 pixels. Nous disposons de cinq réalisations pour



Figure 5. – (a) Exemples de l'ensemble d'apprentissage. (b) Moyenne et premiers vecteurs propres associés au modèle d'apparence de l'œil.

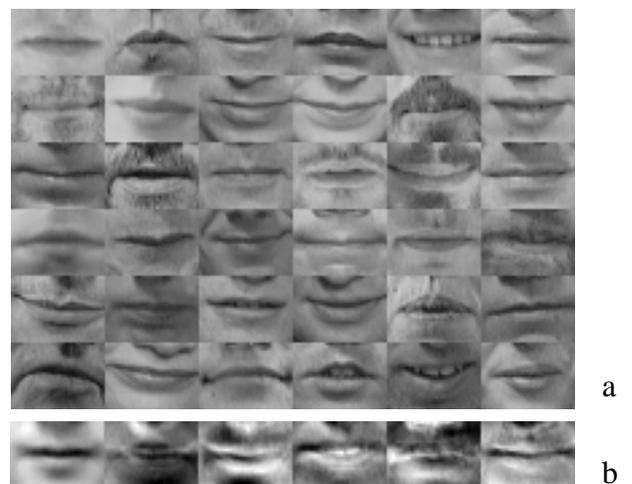


Figure 6. – (a) Exemples de l'ensemble d'apprentissage de la bouche. (b) Moyenne et premiers vecteurs propres associés au modèle d'apparence.

chaque geste et pour chaque personne. Plusieurs milliers d'images de la main (avec des apparences variées) sont ainsi disponibles pour les tests de détection et ont également été utilisées dans une application de reconnaissance des gestes décrite dans (Hamdan *et al.*, 1999 ; Hamdan, 2001).

L'ensemble d'apprentissage pour les images de main a été obtenu en retenant deux réalisations sur les cinq réalisations de chaque geste. Les autres séquences ont été utilisées pour les tests de détection et de reconnaissance. Les segmentations des imageries de la main dans l'ensemble d'apprentissage ont été obtenues avec le même outil interactif, que pour les primitives faciales. La figure 1 présente des exemples d'images de main disponibles dans la base de d'apprentissage, qui illustrent la grande variabilité des apparences représentées. La figure 4 pré-

sente un exemple typique de détection sur une scène contenue dans la base de données.

4.3.2. modèle original de Moghaddam et Pentland

Dans (Moghaddam et Pentland, 1997), Moghaddam *et al.* ont effectué une comparaison de la méthode du maximum de vraisemblance (MV) avec les critères DFFS (distance à l'espace propre) et SSD (distance quadratique avec l'image moyenne de la base). La comparaison a été menée sur la base de données du MIT (plus de 7000 images), dans une tâche de détection de l'œil gauche. Les courbes COR, représentant la probabilité de détection en fonction du taux de fausse alarme, pour différentes valeurs de seuil, montrent clairement la supériorité de l'approche MV par rapport à DFFS ou SSD (Moghaddam et Pentland, 1997). Logiquement le classement obtenu est le suivant : $MV > DFFS > SSD$ (l'estimateur de la SSD est clairement le moins performant, car basé sur un seul représentant de la base d'apprentissage – l'image moyenne – conduisant à un schéma de « template matching » classique). La méthode du MV conduit à un taux de fausse alarme en moyenne 100 fois plus faible que la méthode SSD (Moghaddam et Pentland, 1997).

Comme nous l'avons indiqué plus haut, dans notre implantation, l'estimateur du MV a été testé avec un nombre de vecteurs propres M retenus pour le sous-espace principal E , rendant compte de 50 % de l'information initiale ($\sum_{i=1}^M \lambda_i = 0.5 \sum_{i=1}^N \lambda_i$). Avec ce choix de paramètres, la méthode du MV conduit, sur les bases de visages traitées, à un taux de bonne détection de 96 % (environ 40 fausses détections pour plus de 1000 tests réalisés pour la détection de l'œil gauche et de la bouche), ce qui est cohérent avec les résultats obtenus par Moghaddam et Pentland (Moghaddam et Pentland, 1997) sur la base de données FERET.

L'estimateur du MV fait également preuve d'une certaine robustesse aux occultations partielles et est capable de généralisation (détection de structures faciales pour des personnes n'apparais-

sant pas dans la base d'apprentissage). La figure 7 illustre ces deux points pour un modèle de bouche (Fig. 7a : occultation par une main posée devant la bouche ; Fig. 7b : sujet à moustache, non présent dans l'ensemble d'apprentissage). D'autres exemples sont présentés dans (Hamdan, 2001).

Un taux de bonne détection de 99 % a par ailleurs été constaté pour la segmentation de la main dans les séquences vidéo (ce taux est de 100 % pour les séquences appartenant à l'ensemble d'apprentissage). Ce très bon résultat s'explique en partie par une certaine simplicité des scènes traitées (fond peu texturé, avec peu de structures parasites et pas de leurres - voir figure 4).

4.3.3. comparaison entre le modèle original et le modèle approché

Le modèle original et la version approchée que nous en proposons ont été testés de façon concurrentielle sur les mêmes bases de données, avec le critère du maximum de vraisemblance.

Rappelons que le modèle approché repose sur un calcul séparable (par SVD tronquée) des corrélations entre vecteurs propres et image observée, pour la partie du modèle rendant compte de la distribution des images dans l'espace propre E . La troncature de la SVD conduit à une approximation du noyau de convolution \mathbf{h} , implantant le produit de corrélation, par un noyau \mathbf{h}_k , limité aux $k \ll r$ premières composantes singulières de \mathbf{h} (\mathbf{h}_k est une somme de matrices de rang 1). Le carré de la norme de l'erreur d'approximation entre \mathbf{h} et \mathbf{h}_k est égal à la somme des carrés des valeurs singulières résiduelles (Van-Der-Veen *et al.*, 1993) :

$$\|\mathbf{h} - \mathbf{h}_k\|_F^2 = \sum_{i=k+1}^r \psi_i^2 \quad (9)$$

Nous avons donc tronqué les noyaux de convolution (c'est-à-dire déterminé la valeur de k) en fixant l'erreur relative ϵ_r entre \mathbf{h} et \mathbf{h}_k :

$$\epsilon_r = 100 \frac{\|\mathbf{h} - \mathbf{h}_k\|_F^2}{\|\mathbf{h}\|_F^2} = 100 \frac{\sum_{i=k+1}^r \psi_i^2}{\sum_{i=1}^r \psi_i^2}$$

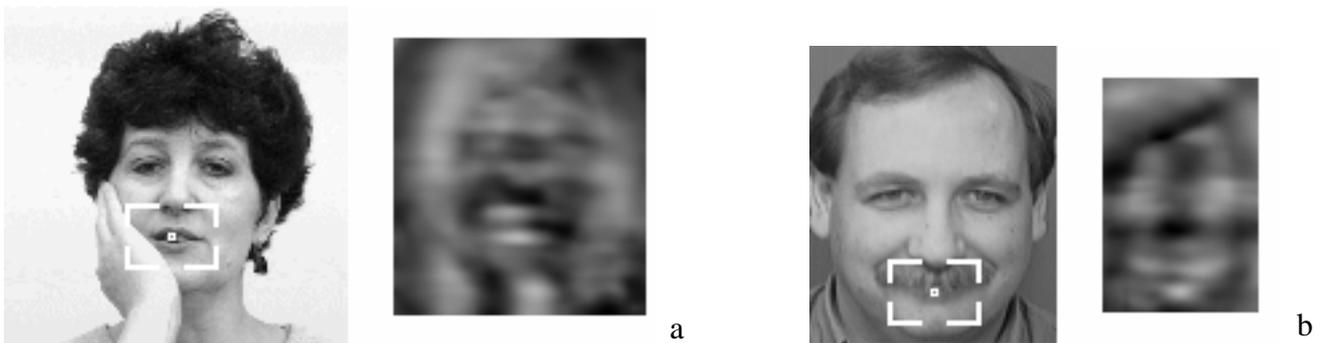


Figure 7. – Robustesse du modèle statistique (modèle de bouche, voir texte).

Aucune différence n'a été constatée entre les deux implantations lorsque l'erreur relative ϵ_r ne dépasse pas 2 ou 3 %. Ce niveau d'erreur permet de ne pas perturber de façon significative la norme et la direction des vecteurs propres, donc la distribution de probabilité associée au modèle. Notons que ce niveau d'erreur relative acceptable (qui peut paraître faible) conduit à retenir un très petit nombre de valeurs singulières pour la représentation par SVD des vecteurs propres (typiquement $k = 3$ ou 4 pour l'image moyenne μ , $k = 5$ à 10 pour les vecteurs propres, ceci pour des matrices de taille 100×100). L'image moyenne μ est plus « lisse » que les vecteurs propres du modèle, elle se prête donc à une réduction d'information plus importante. La réduction de dimension due à la SVD est donc très significative, et fait tout l'intérêt de la méthode. Des gains en temps de calcul supérieurs à 10, ont été observés dans ces tests (voir également tableau 1).

Lorsque l'erreur relative est comprise entre 3 et 10 % la détection en position est simplement biaisée d'un ou deux pixels par rapport au modèle original. Pour des erreurs relatives supérieures, on observe une dégradation importante des performances avec de nombreuses erreurs de détection et fausses alarmes.

5. conclusion

Nous avons proposé dans cet article une modélisation approchée de l'apparence par un modèle statistique gaussien, reposant sur une interprétation probabiliste de l'ACP. La modélisation prend en compte non seulement la distribution des exemples d'apprentissage sur le sous-espace E des premières composantes de l'ACP, mais approche également la distribution de l'ensemble d'apprentissage sur le complément orthogonal, fournissant ainsi une description complète. L'approximation « séparable » du modèle, que nous proposons, s'est montrée équivalente, en termes de performances (détection) au modèle initial, avec un gain en complexité calculatoire significatif (de l'ordre de 10 à 15). Le gain obtenu devrait être encore plus important, dans le cas d'images volumiques : une application concernant la détection/localisation de structures dans des images médicales 3D est en cours d'étude.

Une autre caractéristique intéressante de la méthode d'approximation est qu'elle se prête à une extension immédiate vers une représentation non linéaire, en modélisant des données non gaussiennes par un mélange de lois gaussiennes (Moghaddam et Pentland, 1997). L'approximation proposée peut alors simplement être appliquée à chaque composante du mélange gaussien.

A. annexe : complexité comparée des deux modèles

A.1. complexité du modèle original

Considérons un modèle d'imagette de hauteur et de largeur L , ce qui correspond à un vecteur d'observation de dimension $N = L^2$ pixels. Considérons les opérations nécessaires pour chaque imagette dans la zone de recherche.

L'évaluation directe de la log-vraisemblance du modèle (8) nécessite pour chaque position $p = (i, j)$ de la fenêtre glissante, la normalisation de l'imagette $\mathbf{x}^{(i,j)}$ (centrée en p) en moyenne et en variance, ce qui entraîne N additions-multiplications plus $2N$ additions. Le calcul de la TKL (2), implique ensuite le calcul de $\mathbf{x}^{(i,j)} - \mu$, soit N additions, ainsi que la projection de $\mathbf{x}^{(i,j)} - \mu$ sur les M premiers vecteurs propres, donnant lieu à $M N$ additions-multiplications. L'évaluation du second terme de (8) implique le calcul de la « SSD » en N additions-multiplications. Le terme de normalisation de la gaussienne (8) n'est pas évalué, puisqu'il est identique pour toutes les positions de la fenêtre d'analyse. Pour être précis, il faudrait ajouter $2M$ additions-multiplications plus M multiplications pour le calcul de $\sum_{i=1}^M \frac{y_i^2}{2\lambda_i}$ et $\sum_{i=1}^M y_i^2$. Ce nombre d'opérations est toutefois négligeable dans notre application, où M est de l'ordre de 5 vecteurs propres.

Au total on a donc $(M + 2)N$ multiplications et $(M + 5)N$ additions, pour chaque position p de la fenêtre d'analyse.

A.2. version approchée « séparable »

Normalisation et implantation du modèle séparable. – Pour pouvoir comparer le modèle approché au modèle original, la même normalisation des données doit être adoptée pour le modèle séparable (*i.e.* transformation de chaque imagette $\mathbf{x}^{(i,j)}$ en imagette de moyenne nulle et de variance unité) :

$$\bar{\mathbf{x}}^{(i,j)} = \frac{1}{N} \sum_{l=1}^N \mathbf{x}_l^{(i,j)}$$

$$\sigma^{(i,j)} = \left(\frac{1}{N} \sum_{l=1}^N (\mathbf{x}_l^{(i,j)} - \bar{\mathbf{x}}^{(i,j)})^2 \right)^{\frac{1}{2}}$$

$$\mathbf{x}^{(i,j)} \leftarrow \frac{\mathbf{x}^{(i,j)} - \bar{\mathbf{x}}^{(i,j)} \cdot \mathbf{1}}{\sigma^{(i,j)}}$$

Cependant pour tirer profit de l'avantage calculatoire apporté par l'implantation séparable de la corrélation, les observations

(lignes ou colonnes à traiter), ne doivent pas changer selon la position $p = (i, j)$ de la fenêtre d'analyse. La normalisation locale (sur chaque fenêtre) des imagerie ne permet pas de respecter cette propriété (puisque les moyennes et variances sont calculées de façon glissante selon les équations précédentes, ce qui modifie les données en fonction de la position $p = (i, j)$ de la fenêtre). Le système sous-jacent devient donc non linéaire et non stationnaire.

Pour lever cette difficulté, nous procédons comme suit : considérons la corrélation séparable avec une composante $\mathbf{u} \mathbf{v}^T$ de la SVD (figure 8). Nous effectuons tout d'abord une corrélation des lignes de l'image brute *non normalisée* avec le vecteur ligne \mathbf{v}^T . Les colonnes de l'image intermédiaire ainsi obtenue sont ensuite corrélées avec le vecteur colonne \mathbf{u} .

Nous calculons simultanément et en parallèle les sommes partielles intervenant, en chaque pixel, dans l'évaluation de la moyenne et de la variance associées à une imagerie donnée (calcul des sommes des pixels et des sommes des carrés des pixels). Les résultats de la corrélation avec l'image brute sont normalisés *a posteriori* en utilisant ces sommes partielles. Nous exploitons ici la propriété de distributivité de la convolution par rapport à l'addition, en décomposant la corrélation en :

$$\frac{\mathbf{x}^{(i,j)} - \bar{\mathbf{x}}^{(i,j)} \cdot \mathbf{1}}{\sigma^{(i,j)}} \star \mathbf{h} = \frac{1}{\sigma^{(i,j)}} \left[\underbrace{\mathbf{x}^{(i,j)} \star \mathbf{h}}_{\text{corrélation avec l'image brute}} - \bar{\mathbf{x}}^{(i,j)} \cdot \mathbf{1} \star \mathbf{h} \right] \quad (10)$$

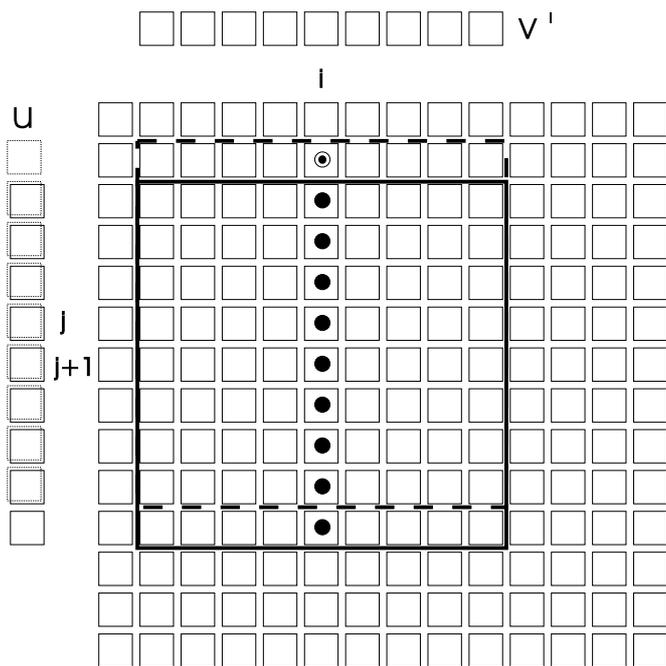


Figure 8. – Corrélation séparable avec une composante de la SVD.

où \mathbf{h} est la réponse impulsionnelle, décomposée par SVD, du filtre représentant la corrélation avec l'un des M vecteurs propres considérés :

$$\mathbf{h} = \sum_{i=1}^k \psi_i \mathbf{u}_i \mathbf{v}_i^T \quad (11)$$

(seules k composantes de la SVD sont conservées sur $L = \sqrt{N}$). Dans l'expression (10), l'écart-type $\sigma^{(i,j)}$ est calculé par la formule suivante, à partir des sommes partielles des pixels avant normalisation et des sommes de leurs carrés :

$$\sigma^{(i,j)} = \left(\frac{1}{N} \sum_{l=1}^N (\mathbf{x}_l^{(i,j)} - \bar{\mathbf{x}}^{(i,j)})^2 \right)^{\frac{1}{2}} = \left(\frac{1}{N} \sum_{l=1}^N (\mathbf{x}_l^{(i,j)})^2 - (\bar{\mathbf{x}}^{(i,j)})^2 \right)^{\frac{1}{2}}$$

Par ailleurs, dans (10), le terme $\mathbf{1} \star \mathbf{h}$ représente la corrélation d'une imagerie composée de 1 avec le filtre \mathbf{h} . Ce calcul peut être mené hors-ligne.

L'évaluation de la vraisemblance (8), implique également le calcul du résiduel de la reconstruction dans l'espace orthogonal, qui, comme nous l'avons déjà indiqué, peut être calculé efficacement à partir de la « SSD » par :

$$\sum_{l=M+1}^N y_l^{(i,j)2} = \|\mathbf{x}^{(i,j)} - \mu\|^2 - \sum_{l=1}^M y_l^{(i,j)2}$$

Le terme « SSD » peut être calculé par :

$$\|\mathbf{x}^{(i,j)} - \mu\|^2 = N - 2 \mathbf{x}^{(i,j)T} \mu + \|\mu\|^2$$

(l'imagerie $\mathbf{x}^{(i,j)}$ étant normalisée en variance à 1, est de norme N). Ceci réduit le calcul à la corrélation séparable de l'image avec la moyenne μ du modèle, ($\|\mu\|^2$ peut être calculé dans une phase préalable hors-ligne).

Soulignons que ces détails d'implantation, quoique techniques, sont importants pour tirer pleinement profit de l'approximation du modèle initial par SVD. Cette façon d'implanter le modèle séparable permet par ailleurs d'assurer qu'il est équivalent au modèle initial, lorsque les SVD de chaque vecteur propre ne sont pas tronquées.

Complexité du modèle séparable. – Considérons maintenant l'évaluation séparable de la log-vraisemblance du modèle (8) pour chaque position $p = (i, j)$ de la fenêtre glissante. On suppose que la SVD de chaque vecteur propre ou de l'image moyenne μ est limitée à k termes. La normalisation de l'imagerie $\mathbf{x}^{(i,j)}$ en moyenne et en variance, réalisée de façon séparable peut être menée en 4 additions-multiplications et 4 additions, en calculant les moyennes partielles de façon glissante, sur le principe d'un terme « entrant » et d'un terme « sortant ». Le calcul de la TKL (2), implique ensuite la projection séparable

de $\mathbf{x}^{(i,j)}$ sur les M premiers vecteurs propres, donnant lieu à $M(2kL)$ additions-multiplications. L'évaluation du second terme implique le calcul de la corrélation séparable avec la moyenne μ du modèle en $2kL$ additions-multiplications. Si l'on néglige le coût de la normalisation, on a donc au total $2k(M+1)L$ multiplications et additions.

Le rapport des complexités du modèle original et du modèle séparable est donc, pour le nombre de multiplications :

$$r_{mult} = \frac{(M+2)L}{2(M+1)k}$$

pour les additions :

$$r_{add} = \frac{(M+5)L}{2(M+1)k}$$

BIBLIOGRAPHIE

- H. Andrews et B. Hunt (1977), *Digital Image Restoration*, Prentice-Hall.
- R. Dahyot, P. Charbonnier et F. Heitz (2001), Détection robuste d'objets : une approche par modèle d'apparence. In *18e Colloque GRETSI*, Toulouse.
- K. Fukunaga (1990), *Introduction to Statistical Pattern Recognition*, Academic Press.
- G. Golub et C.V. Loan (1989), *Matrix Computations*, John Hopkins series in the mathematical sciences, Baltimore, John Hopkins University Press.
- R. Hamdan (janvier 2001), *Détection, suivi et reconnaissance des formes et du mouvement par modèles probabilistes d'apparence*. Thèse de l'Université Strasbourg 1, ftp://picabia.u-strasbg.fr/pub/www/hamdan/these.pdf.
- R. Hamdan, F. Heitz et L. Thoraval (1999), Gesture localization and recognition using probabilistic visual learning, In *IEEE Int. Conf. Computer Vision Pattern Recognition*, Fort Collins, Colorado.
- A. Jain, R. Duin and J. Mao (2000), Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37.
- D. Lowe (1991), Fitting parameterized 3D models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5), 441-450.
- T. Minka (1999), Automatic choice of dimensionality for PCA. Technical Report TR No 514, MIT, Medialab, Vision and Modeling Group.
- B. Moghaddam (1999), Principal manifolds and Bayesian subspaces for visual recognition. Technical Report TR-99-35, Mitsubishi Electric Research Laboratory, http://www.merl.com.
- B. Moghaddam, T. Jebara et A. Pentland (1999), Bayesian modelin of facial similarity, *Advances in Neural Information Processing Systems*, 11.
- B. Moghaddam et A. Pentland (1997), Probabilistic visual learning for object representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 696-710.
- H. Murase et S.K. Nayar (1995), Visual learning and recognition of 3-D objects from appearance, *International Journal of Computer Vision*, 14(1), 5-24.
- W. Murase et M. Lindenbaum (1995), Partial eigenvalue decomposition of large images using spatial temporal adaptive method. *IEEE Transactions on Image Processing*, 4(5), 620-629.
- G. Saporta (1992), *Probabilités, Analyse des données et Statistique*, Technip.
- C. Schmid et R. Mohr (1997), Local grayvalue invariants for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 530-534.
- H. Schneiderman et T. Kanade (2000), Statistical method for 3D object detection applied to faces and cars. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, Hilton Head, South Carolina.
- M.E. Tipping et C.M. Bishop (1997a), Mixtures of principal component analyzers, Technical Report NCRG/97/003, Neural Computing Research Group, Dept. of Computer Science & Applied Mathematics, Aston University, Birmingham, UK.
- M.E. Tipping et C.M. Bishop (1997b), Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Dept. of Computer Science & Applied Mathematics, Aston University, Birmingham, UK.
- M. Turk et A. Pentland (1991), Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, 3(1), 71-86.
- A. Van-Der-Veen, E.F. Deprettere et A.L. Swindlehurst (1993), Subspace-based signal analysis using singular value decomposition, *Proceeding of the IEEE*, 81(9), 1277-1308.

Manuscrit reçu le 23 mars 2001

Modèles probabilistes d'apparence : une représentation approchée

LES AUTEURS

Raouf HAMDAN



Docteur de l'Université Louis-Pasteur (Strasbourg I, 2001), Raouf Hamdan est actuellement Maître de Conférences à l'Université de Damas (Syrie). Titulaire du DEA « Photonique et Image » (Université Louis-Pasteur, 1997), il a effectué sa thèse au LSIIT (Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, UPRES-A CNRS 7005) dans le groupe « Analyse multi-images ».

Ses travaux de recherche ont plus particulièrement porté sur la détection, le suivi et la reconnaissance des formes et du mouvement par modèles probabilistes d'apparence.

Fabrice HEITZ



Ingénieur ENST Bretagne (1984), Docteur Télécom Paris (1988), Fabrice Heitz a été de 1988 à 1994 Chargé de Recherches INRIA à l'IRISA (Rennes) (dans le projet Temis). Il est actuellement Professeur de Traitement du Signal et des Images à l'École Nationale Supérieure de Physique de Strasbourg (ENSPS). Il anime un groupe de recherche en analyse multi-images au sein du LSIIT (Laboratoire des Sciences de l'Image, de

l'Informatique et de la Télédétection, ESA CNRS 7005). Ses domaines d'intérêt incluent la modélisation statistique et les modèles déformables appliqués à l'analyse multi-images et à l'imagerie médicale.

Laurent THORAVAL



Docteur de l'Université de Rennes I en Traitement du Signal et Télécoms (1995), Laurent Thoraval est actuellement Maître de Conférences à l'Université Louis Pasteur - Strasbourg I. Il est membre du groupe de recherche « Analyse multi-images » du LSIIT (Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, ESA CNRS 7005). Ses domaines d'intérêt incluent la modélisation markovienne cachée et

la fusion de données appliquées à l'analyse multi-images et à l'imagerie fonctionnelle cérébrale.