Détection de visages sur des images fixes par combinaison de classifieurs discriminants et de modèles

M.Milgram[1], R.Belaroussi[1] ,L.Prévost[1]

[1]LISIF-PARC, Université Paris VI, 4 place Jussieu 75005 Paris

mauml@ccr.jussieu.fr

Résumé – Cette communication présente une technique de détection de visages sur des images fixes couleur quelconques applicable à l'indexation de séquences vidéos ou à des bases d'images. Les méthodes exposées sont sans segmentation (globale) combinant une étape par modélisation (fusionnant deux classifieurs) et une étape discriminante exploitant la modélisation. On observe une très bonne capacité à détecter des visages d'orientation variée sur un fond quelconque, même avec des résolutions faibles.

Abstract – This paper presents a face detection method for still colour images. It can be applied to indexation of video sequences or for images data bases. Our methods are segmentation free (global) combining a modelization step (merging 2 classifiers) and a discrimination step using the modelization. We have obtained a very good accuracy for face detection for varied orientations and backgrounds, even with low resolution.

1. Introduction

La détection de visage dans une image fixe sans hypothèse particulière est un problème difficile en raison de la très grande variabilité de la forme à détecter (image d'un visage quelconque d'orientation et de taille quelconque avec un éclairement quelconque). Comme pour beaucoup de problèmes de détection, on se trouve dans la quasi-impossibilité de définir la classe adverse, les « non-visages », ce qui incite les chercheurs à s'orienter vers une approche par modèles. Les solutions mises en œuvre dans de nombreuses applications de détection de visage (biométrie, détection de présence, visiophonie, indexation, détection de l'hypovigilance, réalité virtuelle, lecture labiale) commencent par simplifier le problème en ajoutant une ou plusieurs hypothèses: caméra fixe et fond connu (utilisation d'une image de référence sans visage), utilisation du mouvement, hypothèses fortes sur la position, fond propice à l'extraction de la silhouette, maîtrise de l'éclairement (par exemple infra-rouge). La localisation (on sait que le visage est présent mais on ne sait pas où) n'est guère plus simple si on ne fait pas d'hypothèses.

2. Les solutions proposées

On retrouve les deux grandes classes communes à la Reconnaissance des Formes : structurelle et globale. Les approches structurelles [5] cherchent à détecter des primitives du visage (yeux, bouche, nez, contour de la tête) puis à combiner les résultats de ces détections grâce à des modèles géométriques et radiométriques, notamment via des modèles déformables [6]. Les approches globales traitent une vignette de l'image totale en la codant sous la forme d'un vecteur [3] (moments, projection, codage rétinien). Les deux approches utilisent ensuite des bases d'apprentissage et de test pour estimer les paramètres du classifieur. Pour les approches globales, ces paramètres peuvent être des poids (réseau de neurones) ou les termes d'une matrice de covariance (classifieur statistique). Il faut alors choisir entre une approche par modélisation et une approche discriminante. Pour la modélisation, on n'a pas à fournir de contre-exemples ce qui peut sembler un avantage mais qui diminue en fait l'efficacité du classifieur : la généralisation dans un espace de grande dimension (dimension 400 pour des vignettes 20x20) est difficile si on ignore où se trouvent les vecteurs susceptibles d'être confondus.

3. Notre approche : modéliser puis discriminer

Notre approche consiste à utiliser l'étape de modélisation pour extraire une base de contre-exemples pertinents. L'étape de modélisation peut se composer elle-même de plusieurs modules spécialisés, par exemple un module « colorimétrique » et un module « luminance ». Nous montrons dans la section « résultats » la comparaison entre une approche cascade (modélisation+discrimination) et les approches simples (modélisation ou discrimination).

3.1 Couleur chair

La couleur ou teinte chair a souvent été exploitée pour détecter les visages. C'est en effet une caractéristique de bas

niveau (pixel), assez facile à déterminer (en complexité) et quasi-insensible aux facteurs d'échelle (donc ne nécessitant pas de multi-résolution). Néanmoins, il est connu que la couleur est peu fiable dans les zones sombres, sensible à l'éclairage ambiant et bien sûr restrictive (elle exclue l'infra-rouge par exemple).

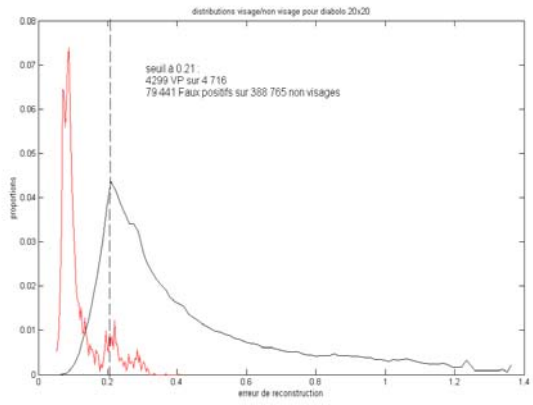
Si les composantes chromatiques habituelles sont (R,G,B), un « bon » vecteur caractéristique dit « chromatique étendu à 6 composantes » est **V**(R,G,B) défini par [1]:

V(R,G,B)= [R,G,B,I, Rg , By] avec:

I = [Log(R) + Log(B) + Log(G)] / 3;Rg = Log(R) – Log(G);By = Log(B) – [Log(G) + Log(R)]

3.2 Réseau auto-associateur

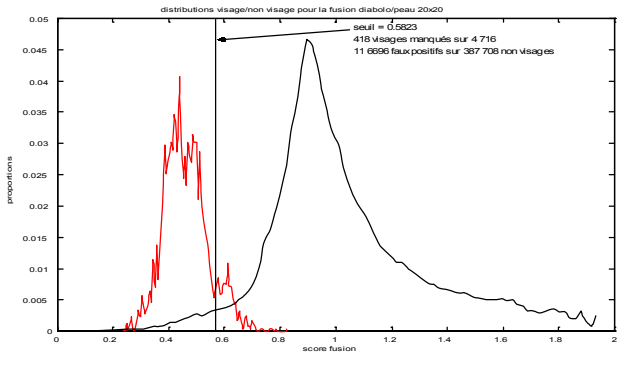
Le traitement de la luminance est réalisé par un réseau de neurones auto-associateur (réseau diabolo). Un tel réseau, que nous avions déjà utilisé pour la reconnaissance de caractères [4], est entraîné à fournir en sortie une image **identique à celle mise en entrée** en réalisant une **compression spécialisée** car la couche cachée comporte un nombre de cellules nettement inférieur à celui de l'entrée ou de la sortie. Une image de non-visage sera en principe mal compressée et donnera une erreur de reconstruction plus importante (voir la figure ci-dessous). Nous avons testé des auto-associateurs ayant des rétines d'entrée (et de sortie) 12x12 ou 20x20 et moins de …neurones cachés. Le nombre de cellules cachées correspond approximativement au nombre de paramètres nécessaires pour représenter les images de la base. Si on suppose que les images forment un nuage de points de l'espace à $\mathbf{R}^{400}$ répartis selon une distribution gaussienne de dimension 400, on peut procéder à une analyse en composantes principales. On constate alors qu'avec les 40 premiers vecteurs propres de la matrice de covariance (de taille 400x400), on récupère 88% de l'inertie du nuage. Néanmoins, si on projette une image quelconque sur ce sous-espace (*eigen face* de Pentland), il apparaît que ce nouveau vecteur dans $\mathbf{R}^{40}$ est nettement insuffisant pour reconnaître un visage. Nous en déduisons que l'hypothèse *gaussienne* est inexacte et que le nuage est en fait un mélange de gaussiennes correspondant notamment aux différentes attitudes et cadrages.



Le réseau auto-associateur et le classifieur de couleur « peau » sont fusionnés par combinaison linéaire [2].

Classement Vérité	V	NV	Total
Visage	4 299	417	4 716
Non-visages	79 441	309 324	388 765

Réseau auto-associateur 20x20, seuil de décision = 0.220



Classement Vérité	V	NV	Total
Visage	4 298	418	4 716
Non-visages	11 169	377 042	388 765

Fusion de l'auto-associateur avec le détecteur de peau, seuil=0.5823

Le nombre de fausses alarmes est divisé par 7.

4. Discrimination

Malgré ses avantages, la modélisation seule ne peut capter tous les traits caractérisant un visage en l'absence complète de non-visages. Nous utilisons ce module pour sélectionner, dans une base de plus de 500 000 vignettes extraites de 682 images sans visage et choisies « au hasard » (paysages,

intérieurs,etc.), les vignettes intéressantes, c'est à dire dont le score (vu par le module modélisation) est proche de celui de visages. Plus précisément, en fixant une probabilité P_d de détection (par exemple : 0.95), on obtient un seuil sur le score issu du module modélisation, seuil qui détermine les vignettes de non-visages intéressantes. En fait nous considérons, comme dans les approches symboliques où la notion de *near miss* est utilisée, que les *non-visages pertinents* sont ceux qui peuvent être confondus avec un visage pour une modélisation donnée. Nous entraînons alors un MLP discriminant sur cette base : visages + non-visages pertinents. Finalement le schéma général est le suivant :

Pour toutes les homothéties(1) de l'image initiale acceptées faire :

1. *sélectionner une vignette Im*
2. *si l'écart-type de Im est inférieur à S1, alors Im est un « NON-VISAGE » aller en 5.*
   - *sinon, calculer la sortie D du réseau auto-associateur et le score P du détecteur de peau si D-α.P < S2 alors Im est un « VISAGE » , aller en 5.*
   - *si D- αP>S3 alors Im est un « NON-VISAGE » , aller en 5,*
3. *sinon :*
4. *M=sortie du discriminant ; M>S4 alors Im est un « VISAGE »,sinon Im est un « NON-VISAGE »*
5. *Passer à la vignette suivante*

(1) : si la rétine est 12x12 et la taille maximum admise pour un visage est 50x50, on applique toutes les homothéties dont les rapports varient de 1 à 12/50 avec un pas suffisant.

Le seuil S1 (écart-type) sert à écarter les zones homogènes et correspond donc au niveau du bruit (10% de la dynamique des niveaux de gris). Le coefficient α a été déterminé par recherche exhaustive sur la base d'apprentissage et vaut 2.5. Les seuils S3 et S2 ont été déterminés à partir des histogrammes des sorties pour les différents modules soit : S2=0.40 et S3=0.80. D'autres schémas de fusion sont bien sûr possibles, notamment en prenant S3 infini ou S2 nul. Le seuil S4 a été pris nul car les sorties du réseau discriminant sont concentrées autour des valeurs extrêmes +1 et –1 malgré un arrêt précoce de l'apprentissage.

Classement / Vérité	V	NV	total
Visage	4653	63	4716
Non-Visage	68	8886	8954

Résultats de l'algorithme ci-dessus

Les non-visages ont été extraits de 103 images à 2 résolutions différentes seulement pour *éviter* une trop *forte redondance des imagettes*

Le taux de détection est maintenant de 98,6% (contre 90% avec la modélisation seule) et le taux de fausse alarme est de 0.7% (contre 3% avec la modélisation seule) ce qui donne un taux d'erreur global de 0.9%. Ces résultats sont cependant délicats à interpréter car, pour une image donnée, il suffit que l'homogénéité ou la couleur soit assez discriminante pour faire chuter le taux de fausse alarme. Nous ne pouvons prétendre avoir épuiser tous les types d'images, notamment texturées. Notons que pour les visages, le taux d'activation du réseau discriminant (zone ambiguë) est de 10.7% et de 13.2% pour les non-visages. Par ailleurs, aucun visage n'est éliminé par le critère d'homogénéité (écart-type>10) mais 45% des non-visages satisfont aussi ce critère. Le MLP traite avec succès 97.9% des visages et 75.2% des non-visages qui lui sont soumis, c'est à dire ayant été ambigus pour le premier étage.

5. Conclusion

Nous avons montré que l'information colorimétrique n'est pas suffisante seule mais permet d'améliorer sensiblement les performances d'un détecteur de visages. Notre processus de sélection des contre-exemples par le module « modélisateur » nous a permis d'entraîner efficacement un réseau MLP discriminant qui obtient d'excellents résultats dans la zone de confusion. Il apparaît néanmoins que le « cadrage » du visage (dans la base d'apprentissage et en test) et l'orientation du visage peuvent créer des problèmes aux modèles d'apparence non structurels comme le nôtre. Une coopération avec des approches « contour » (silhouette du visage ou tête+épaule) ou l'utilisation des symétries ont été utilisées par ailleurs et nous permettrons d'augmenter encore nos performances.

Références

[1] J.P.Kapur, Face detection in color images, *EE499 Design Project, University of Washington Department of Computer Engineering, 1997*

[2] L.Prevost & M. Milgram, Coopération pour la Reconnaissance de Caractères Dynamique Isolés, *RFIA'98, Vol. 3, pp. 233-240, Clermont-Ferrand, Janvier 1998.*

[3] H. Rowley, S. Baluja &T. Kanade, Neural Network-Based Face Detection, *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998.*

[4] H.Schwenk, M.Milgram : Transformation invariant auto-association with application to handwritten character recognition, *NIPS'7 (Neural Inf. Proc. Syst.),pp 991-998, 1995*

[5] G. Yang & T. S. Huang, Human Face Detection in Complex Background, *Pattern Recognition, vol. 27, no. 1, pp. 53-63, 1994.*

[6] A.Yuille, P. Hallinan & D. Cohen, Feature Extraction from Faces Using Deformable Templates, Int J. Computer Vision, vol. 8, no. 2, pp. 99-111, 1992