CORE
Metadata, citation and similar papers at core.ac.uk
Provided by I-Revues

Combinaison de classifieurs pour la localisation de visage

R. BELAROUSSI L. PREVOST M. MILGRAM

[1] Université Pierre & Marie Curie, Laboratoire des Instruments & Systèmes d'Ile de France Groupe PARC

4 Place Jussieu, 75252 Paris Cedex 5, BC252

Résumé – Dans cette communication, nous présentons une méthode de localisation de visages dans des images couleurs combinant trois détecteurs respectivement anthropomorphique (détecteur basé sur un modèle d'apparence neuronal), géométrique (détecteur d'ellipse utilisant une Transformée de Hough Généralisée sur l'image des orientations de gradient) et colorimétrique (détecteur de teinte chair utilisant un seuillage dans l'espace CbCr). La combinaison linéaire de ces trois détecteurs produit une carte de probabilités. La localisation du visage dans l'image correspond au maximum absolu de cette carte. Nous montrons l'apport de la combinaison sur le taux de localisation des détecteurs pris isolément.

Abstract *– In this communication, we present a method to localize faces in color images that combines three detectors: an anthropomorphic detector (using a neural appearance-based model), an ellipse detector (using the generalized Hough transform on gradient orientation image) and a coarse skin color model in CbCr space. Given an input image, we compute a kind of probability map by combining linearly the three detectors output. The face position is then determined as the location of the absolute maximum over this map. Improvement of localization rates of individual detectors is clearly shown and results are very encouraging.*

1. Introduction

La localisation de visages est un domaine de recherche en continuelle expansion en raison de ses nombreuses applications : interaction homme-machine, indexation, biométrie … De nombreuses méthodes ont été développées ces dernières années et sont décrites dans un état de l'art très complet [12]. On peut regrouper ces méthodes selon les deux grandes classes communes à la Reconnaissance des Formes : structurelle et globale. Les approches structurelles [11] cherchent à détecter des éléments caractéristiques 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 [13], ou par l'analyse de "constellations" [1]. Les approches globales traitent une vignette de l'image totale en la codant sous la forme d'un vecteur (codage rétinien, moments, projection) et estiment les paramètres de classifieurs neuronaux [4] ou statistiques [10]. Dans cette communication, nous proposons une nouvelle approche basée sur la combinaison de trois détecteurs respectivement anthropomorphique, géométrique et colorimétrique ; combinaison qui améliore clairement les performances de chacun des détecteurs considéré séparément.

2. Les détecteurs de base

Une partie de l'information contenue dans l'image d'un visage se trouve dans l'orientation du gradient des contours. Le premier avantage des orientations des contours est leur relative invariance à la couleur de la peau. Le second est lié à la présence de parties concaves (yeux, bouche) et convexes (nez) dans un visage qui créent des contours. Deux des trois détecteurs présentés dans cette section utilisent cette information : le modèle d'apparence (réseau de neurones Diabolo) et le modèle d'ellipse (Transformée de Hough Généralisée).

Notre technique, parfaitement classique pour déterminer les orientations, est la suivante :

- convolution de l'image par un masque circulaire dont le diamètre est de 10% de la largeur de l'image,
- estimation du champ de gradient (I_x, I_y) avec un masque de Roberts,
- seuillage des modules du gradient afin de déterminer les pixels de contours,
- calcul en ces points de l'orientation discrétisée sur N=36 valeurs.

Pour le Diabolo le seuil est défini pour chaque vignette 13x17, à 20% des modules les plus grands (on conserve donc 20% de pixels comme étant de contour). Pour la transformation de Hough, le seuil est appliqué sur l'image entière. Il est de 12 et a été optimisé sur 168 images d'apprentissage.

2.1 Le réseau diabolo : un modèle d'apparence du visage

Chaque pixel de contour de la vignette se voit attribuer deux caractéristiques (I_{cos},I_{sin}) :

$$I_{\cos}(i,j)=\cos\left(\frac{2\pi}{N}\cdot orienx(i,j)\right) \text{ et } I_{\sin}(i,j)=\sin\left(\frac{2\pi}{N}\cdot orienx(i,j)\right)$$

avec $orienx = round\left(\frac{N}{2\pi}\arctan\frac{I_y}{I_x}\right)$ mod(N). Les pixels

hors contours reçoivent les valeurs (0,0). Cette étape produit deux tableaux I_{cos} et I_{sin} que l'on réduit à la taille 13x17 par interpolation bi-cubique. On extrait les valeurs résultantes se trouvant à l'intérieur d'un masque elliptique afin d'éliminer les effets des bords de l'imagette. Un vecteur de 290 éléments est ainsi obtenu (on notera la réduction de dimension comparée à 13x17x2=442).

Le traitement de ces exemples est réalisé par un réseau de neurones auto-associateur (réseau Diabolo [3]). Ce réseau 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 (figure 1). 1 602 images de visage sont utilisées en apprentissage et 178 en validation croisée. Les poids et biais du réseau de neurones sont estimés par descente de gradient de l'erreur quadratique avec adaptation du pas d'apprentissage. Une recherche exhaustive a montré que pour 290 entrées (rétine de taille 13x17), une couche cachée de 18 neurones est optimale. L'image en niveau de gris est parcourue, à l'échelle correspondant à la taille du visage, par une rétine 13x17, et à chaque position de l'image une erreur de reconstruction est calculée. Une carte des erreurs de reconstruction est ainsi obtenue, nous l'appelerons *DiaboloMap* dans la suite.

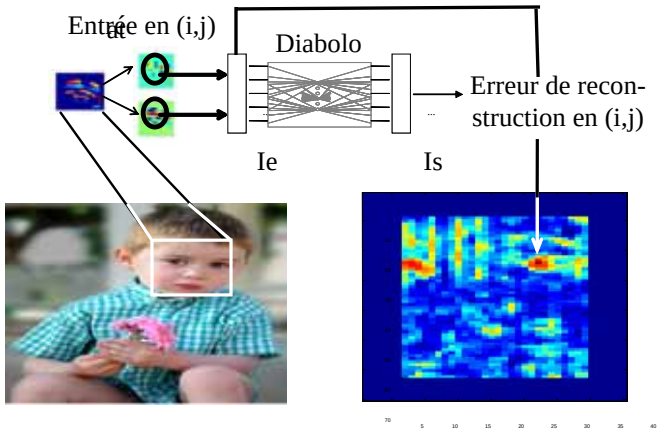

FIG. 1 : Erreur de reconstruction du diabolo (les couleurs chaudes correspondent aux erreurs de reconstruction faibles).

2.2 Détecteur d'ellipse basé sur la Transformation de Hough Généralisée

Une Transformation de Hough Généralisée (THG) est réalisée sur la carte des orientations. Concernant la détection d'ellipses, il existe une structure simplifiée de la THG basée sur les propriétés géométriques des ellipses. La méthode consiste à accumuler les votes d'une demi-droite ayant pour point de départ un pixel de contour et dont la direction et le sens sont donnés par l'orientation du contour en ce point. L'orientation de l'ellipse est supposée connue : les visages sont modélisés par une ellipse d'axe vertical d'excentricité donnée. Alors, pour chaque point de contour *M*, une simple table précalculée spécifie l'angle entre la tangente au contour *Mt* et le rayon *MO* (*O* étant le centre de l'ellipse passant par *M*).

Il en résulte un tableau de vote dont le maximum correspond à la position dans l'image du point le plus susceptible d'être le centre de l'ellipse. Comme le fond est souvent structuré et complexe, ce maximum ne localise qu'approximativement la position du visage. Afin de diminuer l'effet de ce fond sur le tableau de vote, ce dernier est d'abord lissé puis parcouru par un "chapeau mexicain" de taille 13x17. La somme des éléments de chaque sous-tableau, pondérés par les coefficients du masque, donne un nouveau score. La carte de l'image qui en résulte (figure 2) sera nommée *HoughMap* dans la suite de cet exposé.

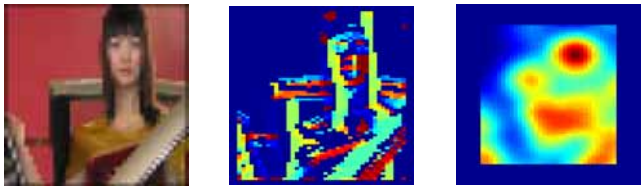

FIG. 2 : image originale, orientation du gradient des contours et tableau de vote.

2.3 Modèle de teinte chair

Indépendamment de l'information d'orientation des contours, une détection de peau basique est implémentée dans l'espace *YCbCr* [2]. L'avantage de l'espace colorimétrique *YCbCr* est qu'il sépare l'information de luminance (*Y*) de celle de la chrominance (*Cb*-*Cr*). Notre filtre de teinte chair est défini par les intervalles suivants : $Cb \in [105 \quad 130]$ et $Cr \in [135 \quad 160]$. La peau étant caractérisée par des informations de chrominance spécifiques, ce filtre peut s'appliquer aux différents types de couleur de peau rencontrés en fonction des origines ethniques. Ces seuils ont été déterminés expérimentalement en utilisant des images (hors test) contenant des visages. On présente ci-dessous la matrice de confusion de cette opération de seuillage sur un échantillon de 145 736 580 pixels de couleur peau (set 3 de la base ECU [8]) et 667 359 498 pixels non-peau :

TAB. 1 : Matrice de confusion du modèle de couleur peau.

Classification / Etiquette	Peau	Non-peau
Peau	76%	24%
Non-peau	19%	81%

Ce modèle de couleur peau est grossier, et dans certains cas tous les pixels de peau sont masqués, mais la fusion des trois détecteurs nous permet l'utilisation d'un modèle simple. Les seuils utilisés ne sont pas universels du fait que la chrominance dépend dans une certaine mesure de la valeur de la luminance Y [5]. Dans des conditions d'illumination faible ou saturée, les composantes filtrées laissent apparaître des trous. Le masque des pixels "non-peau" bruité ainsi obtenu est alors lissé, et des opérations de morphologie mathématique sont appliquées pour remplir les éventuels trous, et le résultat est binarisé. A chaque position de l'image, le pourcentage de pixel filtré dans une fenêtre 13x17 est calculé. Nous appellerons la carte qui en résulte *SkinMap*.

3. Combinaison des différents détecteurs

Nous avons défini trois détecteurs qui produisent, pour une image couleur donnée, en trois cartes : *DiaboloMap*, *HoughMap*, et *SkinMap*. La combinaison linéaire de ces trois réponses (figure 3) permet d'améliorer les taux de localisation de chaque détecteur pris isolément [9]. La carte résultante est appelée *FusionMap*.

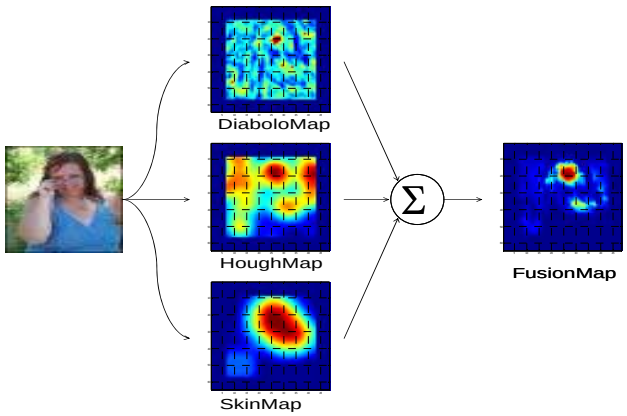

FIG. 3 : Système de localisation du visage

Dans cette optique, la réponse de chaque détecteur est ajustée linéairement dans l'intervalle $[-1\ \ 1]$. Appelant ces cartes normalisées *D*,*H* et *S*, l'imagette 13x17 à la position (i,j) dans l'image originale est alors caractérisée par $I_{i,j} = \left[H_{i,j} \quad D_{i,j} \quad S_{i,j} \right]$. Cent images de la base d'apprentissage ont été utilisées pour apprendre les coefficients de la combinaison linéaire par descente du gradient arrêtée par validation croisée :

$$FusionMap_{i,j} = a \cdot H_{i,j} - b \cdot D_{i,j} + c \cdot S_{i,j} + d$$

avec a=0. 2280, b=0.2620 , c=0. 1229 et d= -0.7198

On notera la faiblesse du poids du détecteur de couleur peau comparé à celui autres détecteurs.

4. Evaluation des performances

Pour déterminer la position du visage, l'image est parcourue par les trois détecteurs précédemment décrits, et le visage correspond à la position du maximum de la carte FusionMap résultante. Nos expérimentations sont réalisées sur la base ECU [8] qui se divise en images en couleur de personnes (set 1), la vérité terrain des visages (set 2) ainsi que la vérité terrain de la teinte chair (set 3). La tâche de localisation sur cette base est particulièrement complexe en raison de la grande variabilité apparaissant dans les images de visages (figure 4) au niveau des caractéristiques intrinsèques (origine ethnique, âge, orientation) et extrinsèques (taille dans l'image, résolution).

Un premier test est réalisé sur 1 353 images (distinctes des images des corpus d'apprentissage et de validation croisée) contenant un seul visage, permettant une évaluation du taux de localisation (nombre de visages correctement localisés divisé par le nombre total de visages). Pour chaque image, la taille du visage est supposée connue, ce qui nous permet d'appliquer une fenêtre glissante de taille fixe à l'image. Cette connaissance revient à celle de la distance du visage à la caméra. Les images en test sont donc réduites de manière à ce que le visage puisse être contenue dans une fenêtre 13x17. Cette taille permet de respecter le rapport hauteur sur largeur d'un visage, et est un bon compromis entre visibilité (pour un œil humain) des caractéristiques d'un visage et l'effort computationnel. Après combinaison, 1 166 visages sur les 1 353 que contient la base de test sont correctement localisés, soit un taux de localisation de 86%. La fusion des trois détecteurs permet ainsi de diminuer le taux d'erreur de 50% (tableau 2).

FIG. 4 : Exemples de localisation.

TAB. 2 Performances des détecteurs et combinaison.

détecteur	diabolo	hough	fusion
Taux de localisation	48.5 %	67 %	**86 %**

Un second test est réalisé sur 205 images (différentes de celle utilisée en apprentissage ou en validation croisée) contenant un total de 482 visages. Dans les images "mono-visage" la position du visage est définie par celle du maximum de FusionMap. Dans une image contenant N visages (N étant supposée connu pour un problème de localisation) les N plus grand maximum locaux (suffisamment distants pour éviter les recouvrements) définissent la position des visages. 396 visages sont correctement détectés (figure 5) sur 482 (82%).

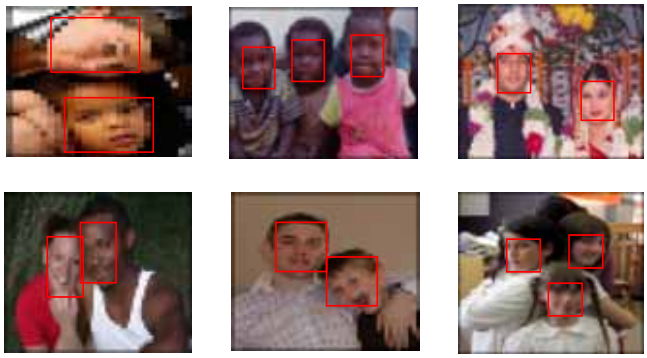

FIG. 5 : Localisation multi-visage.

5. Conclusion et perspectives

Cette communication a pour but de présenter une contribution significative au problème de la localisation de visage. Nous avons présenté trois détecteurs représentant des sources d'information différentes : couleur de la peau, apparence et forme géométrique du visage. Nous avons montré qu'une fusion linéaire de ces trois sources améliore clairement les performances de chacun des détecteurs.

De nombreuses améliorations sont à l'étude. Pour la teinte chair d'abord, nous utilisons un modèle probabiliste avec mélange de gaussiennes [6]. Pour la combinaison de détecteurs, un expert de fusion neuronale devrait aussi améliorer les résultats comparé à la fusion linéaire. De plus un algorithme de détection robuste des yeux a été développé [7] et sera utilisé pour la validation des visages candidats.

Références

[1] Bileschi S.M. & Heisele B., Advances in Component Based Face Detection, *IEEE Int Workshop on Analysis and Modeling of Face and Gestures, 2003.*

[2] Chai D. & Nang K.N., Locating facial region of a head-and-shoulders color image, *Int Conf on Automatic Face and Gesture Recognition, pp.124-129, 1998.*

[3] Féraud R., Bernier O., Viallet J. & Collobert M., A Fast and Accurate Face Detector Based on Neural Networks, *IEEE Trans. PAMI, 23(1), pp. 42-53, 2002.*

[4] Garcia C. & Delakis M., Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection *IEEE Trans. PAMI, 26(11), pp. 1408-1422, 2004.*

[5] Hu M., Worrall S., Sadka A.H. & Kondoz A.M., Automatic scalable face model design for 2D model-based video coding, *Signal Processing: Image Communication, 19, pp. 421-436, 2004.*

[6] Milgram M., Prevost L. & Belaroussi R., Une nouvelle transformation pour la localisation des yeux dans une image de visage monochrome, à paraître, *GRETSI 2005.*

[7] McKenna S.J., Gong S. & Raja Y., Modelling facial colour and identity with gaussian mixtures, *Pattern Recognition 31(12), pp.1883-1892, 1998.*

[8] Phung S.L., Bouzerdoum A. & Chai D., Skin segmentation using color pixel classification: Analysis and comparison, *IEEE Trans. PAMI, 27(1), pp 148-154, 2005.*

[9] Prevost L. & Milgram M., Automatic Allograph Selection and Multiple Expert Classification for Totally Unconstrained Handwritten Character Recognition, *IEEE Int Conf on Pattern Recognition, (1), pp 381-383, 1998.*

[10] Sung K.K & Poggio T., Example-based learning for view-based human face detection, *IEEE Trans. PAMI, 20(1), pp 39-51, 1995.*

[11] Yang G. & Huang T. S., Human Face Detection in Complex Background, *Pattern Recognition, 27(1), pp. 53-63, 1994.*

[12] Yang M.H., Kriegman D. & Ahuja N., Detecting Faces in Images: A Survey, *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002.*

[13] Yuille A., Hallinan P. & Cohen D., Feature Extraction from Faces Using Deformable Templates, *Int Journal Computer Vision, 8(2), pp. 99-111, 1992.*