Détection de visage adaptative

Renaud Séguier
IETR (Institut d'Électronique et de Télécommunication de Rennes)
Équipe Automatique et Communication
SUPÉLEC, Avenue de la Boulaie, 35511 Cesson-Sévigné, France
Renaud.Seguier@supelec.fr

Résumé

Nous présentons dans cet article un système complet permettant de détecter des visages très rapidement dans des conditions difficiles (mauvaise calibration des ccd, environnement extérieur bruité) pour les systèmes embarqués de type mobiles, PDA et consoles de jeux.
Le détecteur exploite sa capacité à être exécuté en temps réel pour fournir une note fiable qualifiant la localisation des visages au cours de la séquence. Cette note permet de focaliser l'attention du détecteur sur la teinte, la forme et le mouvement des visages. Aucun a priori n'est fait quand à la teinte des visages, ce qui n'est pas le cas des détecteurs de visages rapides habituels de ce fait inadaptés à notre contexte applicatif.
Cet algorithme est peu consommateur en temps de calcul (5ms sur un PC courant) et laisse ainsi le temps à d'autres processus d'être exécuté (analyse de visage, clonage, compression, reconnaissance de gestuelle). Sa généricité (code C non optimisé) est illustrée par une implémentation sur PALM (Sony Clié, NX70-V, PalmOS5) et Pentium 4 (Windows 2000).

Mots clefs

Détection de visage, Interactions Homme-Machine, Traitement d'Image Temps-Réel.

1 Introduction

Les systèmes actuels de détection de visage peuvent être classés selon que l'on se base sur le visage entier ou sur des traits caractéristiques du visage [1], [2]. Dans la première approche on génère une base d'exemples à partir de laquelle un classifieur va apprendre ce qu'est un visage (Neural Networks, Support Vector Machine, Principal Component Analysis - Eigenfaces...). Ces systèmes sont très performants [3], [4] mais très lents car lourds à mettre en œuvre. Pour atteindre le temps réel ($\sim$40ms) ils nécessitent une implémentation spécifique (ZISCs, FPGA ou DSP) [5].
Dans la seconde approche on peut distinguer trois niveaux d'analyse. Au niveau le plus rudimentaire, on prend en compte le mouvement, la couleur ou les niveaux de gris pour détecter des régions ("*blobs*") ressemblant à un visage, de face généralement. Cette approche est réalisable en temps réel mais peu robuste. A un niveau d'analyse intermédiaire, on cherche à détecter des caractéristiques indépendantes des conditions lumineuses et de l'orientation des visages. Enfin, à un haut niveau d'analyse, on recherche des traits caractéristiques du visage tels que les yeux, les contours extérieurs, le nez et la bouche que l'on associe à des configurations (*"templates"*) connues a priori ou apprises [6]. Des modèles déformables, des snakes ou des Point Distributed Models (PDM) peuvent être alors utilisés. Ces derniers modèles requièrent une bonne résolution de l'image et sont difficilement réalisables en temps réel. Cependant, une fois le visage détecté, il est alors possible de le suivre en temps-réel [7].

Le travail présenté ici s'insère dans la seconde approche, à un niveau intermédiaire d'analyse. Les contours extérieurs d'un visage sont modélisés par une ellipse parfois en mouvement englobant une zone de teinte particulière.

Dans la très grande majorité de cas, lorsqu'un système de détection est exécuté en temps réel et utilise la teinte chair, celle-ci est connue a priori [8], [6], [9], [10]. Notons que plusieurs équipes [11], [12] mettent en œuvre une adaptation de la teinte du visage, mais partent du postulat que la caméra est correctement calibrée et utilisent donc une signature de la teinte chair connue a priori, signature affinée au fur et à mesure de la séquence.

Dans la classe des détecteurs temps-réel, ce qui fait l'originalité de notre système est sa capacité à localiser des visages de profil ou de face sans pour autant considérer la teinte a priori de la peau, teinte qu'il apprend au fur et à mesure du traitement. Cela le rend robuste aux conditions lumineuses et au calibrage de la caméra, robustesse indispensable dans le contexte des systèmes embarqués (téléphone mobile etc.) où il n'est pas possible de postuler a priori de la teinte des visages [13]. Sa rapidité d'exécution (5ms) rend possible l'exécution d'autres processus (reconnaissance de gestuelle, compression etc.).

Dans la section 2 après avoir présenté le système global, nous détaillerons un à un les différents modules. Dans la section suivante quelques aspects de l'implémentation seront discutés. Dans la section 4 les performances du système seront illustrées sur des séquences réelles et nous conclurons dans une dernière section en discutant des travaux futurs.

2 Système de détection

Nous modélisons un visage comme une ellipse, parfois en mouvement, délimitant une région de teinte particulière que nous apprenons dans les toutes premières secondes et dont le mouvement dans l'image n'est pas chaotique.
L'idée générale est de profiter de la redondance des informations au cours du temps (c'est toujours le même visage qui est détecté plusieurs secondes de suite) pour adapter les paramètres du système. La qualité de la détection va donc en s'accroissant au cours du temps.
Les détections sont qualifiées à l'aide d'une note. Lorsqu'elle est bonne (caractéristiques de forme et de mouvement du visage cohérent au cours du temps) le système focalise son attention en adaptant ses paramètres à la teinte et à la forme du visage détecté.
Le traitement est constitué de huit modules organisés selon le schéma de la figure 1.

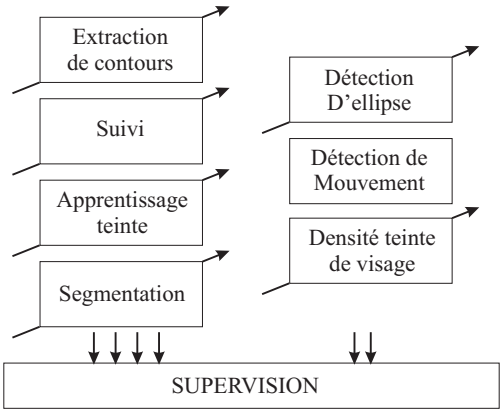

Figure 1 – *Le système de détection*

2.1 Modules de traitement

Les algoritmes utilisés dans chacun de ces modules sont classiques ; pour cette raison ils seront décrits succintement. Les quatre premiers modules (extraction de contours, suivi, apprentissage de la teinte du visage et segmentation) travaillent directement sur les pixels. Un filtre de Shen-Castan [14] extrait les contours de l'image. Le visage détecté à l'instant $t-1$ est recherché à l'instant courant t par un simple block matching. Lorsque le système a confiance dans la détection courante, la teinte du visage est évaluée en analysant les histogrammes des composantes de chrominance $CbCr$ de la zone détectée comme un visage. Enfin, lorsque le système a appris la signature du visage, l'image est segmentée : en chaque pixel une valeur binaire caractérise la présence d'un pixel de teinte chair.
Les trois modules suivants (détection d'ellipse et de mouvement, densité de teinte chair) utilisent les sorties des modules précédents. Une Transformée de Hough Floue Généralisée (THFG) [15] localise les objets de l'image ressemblant à une ellipse. Les contours extraits aux instants précédent et courant sont comparés pour mesurer la quantité de mouvement dans l'image. Dans la zone détectée comme un visage, si le module de segmentation est actif, le pourcentage de pixels caractérisant la teinte du visage détecté est mesuré.

2.2 Module de supervision

Le dernier module est chargé de superviser l'exécution des modules précédemment décrits et de modifier leur paramètres. Si du mouvement a été détecté, il impose au module de détection d'ellipse de ne prendre en compte que les contours qui ont bougés. L'intérêt est qu'alors les contours associés au fond de l'image ne sont pas considérés, rendant plus robuste la détection. Ainsi, la stratégie est la suivante : si les modules de tracking et de détection d'ellipses donnent sensiblement les mêmes valeurs et que du mouvement a été détecté, nous estimons avoir accroché le visage. Si le visage est accroché cinq fois de suite, alors nous avons confiance dans la détection en cours.
Lorsque nous avons confiance dans la détection, le module d'apprentissage de la teinte de la peau est activé. En début de séquence, nous attendons d'avoir analysé cette teinte une dizaine de fois avant de donner la signature de la teinte du visage au module de segmentation. Une fois que la teinte du visage a été apprise, le module qui calcule la densité de teinte chair nous donne pour chaque localisation possible d'ellipse (donnée par la THFG) le taux de pixels caractérisant la teinte du visage précédemment détecté dans l'ellipse considérée. C'est l'ellipse qui affichera le taux maximum qui caractérisera le visage détecté par le système.

3 Implémentation

Afin d'accélérer les traitements, nous exécutons chacun des modules dans un mode multirésolution. Les résultats suivants sont basés sur des images acquises avec une résolution de 320x240 pixels par une webcam usuelle en 4/2/0 $YCbCr$. Notre système est capable de détecter des visages allant de 40 à 100 pixels de large (3 niveaux de résolutions, deux accumulateurs de Hough par résolution).
Ce système a été implémenté sur un Pentium 4, 2.6GHZ de bureautique sans aucune optimisation spécifique (ni matérielle, ni soft). Une Webcam usuelle a été utilisée et permet d'acquérir le signal à 25Hz via le port USB. Notre détecteur s'exécute en seulement 5ms laissant ainsi 35ms par image pour effectuer d'autres traitements (compression, analyse de visage et de gestuelle, clonage). Ces performances sont à comparer à celles de [6] qui met 33ms pour détecter un visage dans une image de 320x240 pixels (résultats extrapolés par rapport à ceux présentés dans l'article : 110ms pour une image sur un P3 800MHZ) ; de [10] qui met 92ms (résultats extrapolés, dans l'article : 300ms pour une image 320x240 sur un P3 800MHZ) et de [13] qui met 66ms pour localiser un visage sur un DSP ;
A notre connaissance, les résultats les plus performants en terme de rapidité et d'efficacité sont ceux de [16] qui se base uniquement sur des contours orientés pour la détection

de visages de face. Ils affichent une robustesse comparable à celles des architectures neuronale et un temps d'exécution très faible (40ms pour la détection de visage de 27x32 pixels dans des images de 320x240 sur un Athlon 1GHZ). Porté sur notre P4 2.6GHZ, leur détecteur atteindrait les 15ms. Nous restons donc trois fois plus rapide tout en étant capable de détecter des visages de profil et de face et en fournissant en plus la teinte des visages détectés.
Nous avons également implémenté le détecteur sur un PALM (Sony Clié, NX70-V PalmOS5) doté d'un ARM cadencé à 200MHZ. Les drivers Sony de la caméra étant indisponibles, il n'est pas possible de traiter le flot vidéo en temps réel. Le détecteur fonctionne donc sur des images fixes acquises au préalable et fortement compressées. Dans ces conditions difficiles (dues à la dégradation des contours apportées par la compression Jpeg), nous parvenons à localiser les visages uniquement lorsque le fond est moyennement perturbé puisque nous n'utilisons ni l'information de teinte chair ni le mouvement des contours orientés.

4 Résultats

Comme on peut le voir sur la figure 2 le système est capable de localiser des visages de tailles variables, en dépit de leur orientation, de leur forme et de la présence de bruit (en terme de contours et de teinte) dans le fond de l'image. Sa capacité à faire la distinction entre des avants-bras et le visage lui-même le rend concurrentiel par rapport aux systèmes de détection temps réel qui se basent sur la teinte chair uniquement.

Figure 2 – *Exemples de détection*

Pour illustrer la capacité de notre système à s'adapter à des teintes de visage différentes, nous avons contraint la caméra à filmer la même scène (illuminée par des néons) en spécifiant une balance automatique des blancs (séquence 1), une lumière intérieur différente du néon (séquence 2, dominante bleu) et une lumière extérieur (séquence trois, dominante jaune). Les histogrammes Cr et Cb réels du visage contenu dans la scène varient de façon importante d'une scène à l'autre (voir figure 3). Les gaussiennes représentée sur cette figure sont celles qui ont été apprises par le système de façon automatique à la fin de l'étape d'adaptation. Même si elles ne sont pas toujours parfaites (à titre d'exemple : la signature Cr de la séquence 2), elles apportent une information utile quand à la chrominance des pixels contenus dans le visage.

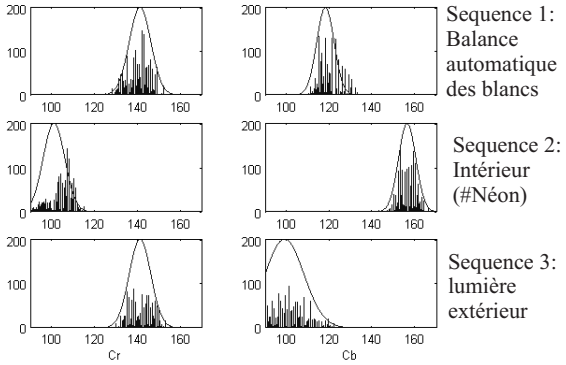

Figure 3 – *Trois signatures différentes de teinte chair*

Nous avons réalisé l'acquisition d'une scène particulièrement contraignante pour notre détecteur : le fond contient beaucoup de contours verticaux et horizontaux qui rajoutent un bruit conséquent dans les accumulateurs de Hough, et nous avons mis des panneaux de teinte chair pour perturber la segmentation (voir figure 5).

Le système met 122 images sur 350 pour apprendre la teinte du visage. Durant cette période de calibration, il localise mal le visage (centre détecté à une distance de plus de 10 pixels du centre du visage) dans 38% des cas.

Afin de mesurer l'apport quantitatif de la segmentation nous avons comparé les erreurs de détection après calibration entre un système n'utilisant pas (figure 4a) ou utilisant (figure 4b) la segmentation en supprimant l'utilisation des contours en mouvement. Le pourcentage d'erreur passe alors de 32% a 3.7%.

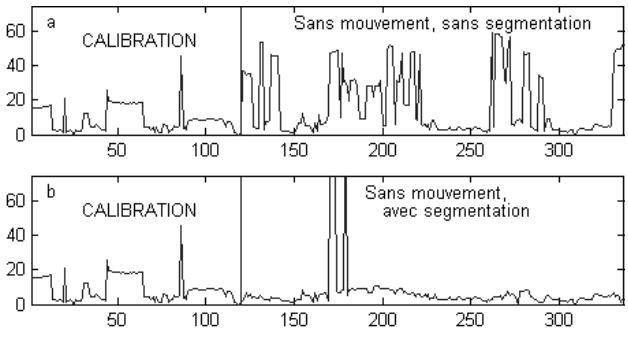

Figure 4 – *Erreurs de localisation au cours du temps*

Après calibration, le détecteur complet (utilisant le mouvement et la segmentation) localise correctement le visage même lorsqu'il se trouve prèt d'un fond de même teinte (dernière image segmentée de la figure 5).

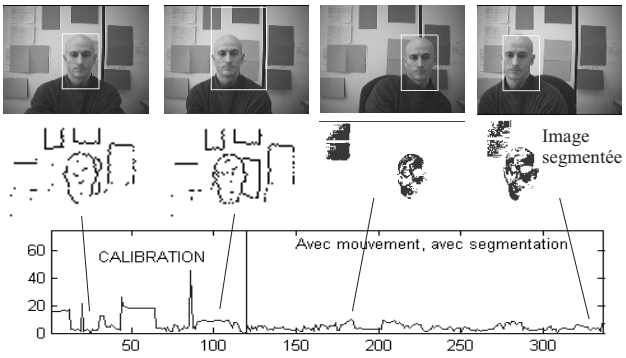

Figure 5 – *Performances du détecteur avant et après calibration*

5 Conclusion

Nous avons présenté un système de détection de visage basé sur les contours orientés, la couleur et le mouvement. Sa particularité repose sur sa capacité à être exécuté en temps réel en dépit du fait qu'aucun a priori n'est fait sur la teinte des visages. Cette teinte est apprise au cours du temps et conduit à une détection de plus en plus robuste. Son très faible temps d'exécution (5ms) le rend attractif dans le contexte des systèmes embarqués (mobiles, PDA, console de jeux) nécessitant une interaction avec l'utilisateur, puisqu'il laisse du temps de traitement à d'autres processus (analyse de visage, clonage, compression, reconnaissance de gestuelle). Notons que notre algorithme fournit la teinte du visage détecté, information pouvant être utilisée pour localiser les mains et donc analyser la gestuelle d'un individu.

Cependant notre détecteur ne parvient pas à localiser correctement les visages lorsque d'autres objets de teinte chair en mouvement et ressemblant parfois à des ellipses (mains et avant-bras) apparaissent dans le champ de la caméra. C'est évidemment un problème de taille dans les applications analysant la gestuelle des personnes. Il nous reste donc à mieux caractériser les visages en enrichissant le modèle détecté par la présence des yeux et de la bouche. Par ailleurs des tests sont en cours sur la base européenne M2VTS [17] afin de comparer les performances de notre algorithme aux systèmes équivalents [16]. Cette comparaison ne se fera que sur des images fixes.

Références

[1] M.H Yang, D. J. Kriegman, et N. Ahuja. Detecting faces in images : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[2] E. Hjelmas et B. K. Low. Face detection : A survey. *Computer Vision and Image Understanding*, 2001.

[3] Raphael Féraud, Olivier J. Bernier, Jean-Emmanuel Viallet, et Michel Collobert. A fast and accurate face detector based on neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[4] C.Garcia et M.Delakis. A neural architecture for fast and robust face detection. Dans *Proceedings of the IEEE-IAPR International Conference on Pattern Recognition (ICPR'02)*, 2002.

[5] F. Yang et M. Paindavoine. Implementation of an rbf neural network on embedded systems : real-time face tracking and identity verification. *IEEE Transactions on Neural Networks*, 2003.

[6] C.C. Chiang, W.N. Tai, M.T. Yang, Y.T. Huang, et C.J. Huang. A novel method for detecting lips, eyes and faces in real time. *Real-Time Imaging*, 9, 2003.

[7] K. Toennies, F. Behrens, et M. Aurhammer. Feasibility of hough-transfrm-based iris localisation for real-time-application. Dans *International Conference on Pattern Recognition*, 2002.

[8] M.J. Chen, M.C. Chi, c.T. Hsu, et J.W. Chen. Roi video coding based on h.263+ with robust skin-color detection tehnique. *IEEE Transactions on Consumer Electronics*, 49(3), 2003.

[9] G.L. Foresti, C. Micheloni, L. Snidaro, et C. Marchiol. Face detection for visual surveillance. 2003.

[10] X. He, Z.M. Liu, et J.L. Zhou. Real-time human face detection in color image. 2003.

[11] V.Girondel, L.Bonnaud, et A.Caplier. Hands detection and tracking for interactive multimedia applications. Dans *International Conference on Computer Vision and Graphics*, 2002.

[12] D. Comaniciu, F. Berton, et V Ramesh. Adaptive resolution system for distributed surveillance. *Real-Time Imaging*, 8, 2002.

[13] K. Imagawa et al.. Real-time face detection with mpeg4 codec lsi for a mobile multimedia terminal. Dans *International Conference on Consumer Electronics*, 2003.

[14] J. Shen et S. Castan. An optial linear operator for step edge detection. *Graphical models and image processing CVGIP*, 54(2), 1991.

[15] R. Séguier, A. Le Glaunec, et B. Loriferne. Human faces detection and tracking in video sequence. Dans *Proc. 7th Portuguese Conf. on Pattern Recognition*, 1995.

[16] B. Froba et C. Kublbeck. Robust face detection at video frame rate on edge orientation features. Dans *International Conference on Automatic Face and Gesture Recognition*, 2002.

[17] S. Pigeon. M2vts. Dans *www.tele.ucl.ac.be/ PROJECTS/M2VTS/m2fdb.html*, 1996.