Détection et suivi robuste de visages en temps réel

O. Bernier M. Collobert D. Collobert

France Télécom R&D DTL/TIC/TNT, Technopole Anticipa, 2, Av. Pierre Marzin, 22307 LANNION Cedex

olivier.bernier@francetelecom.com

Résumé

Suivre le visage d'une personne par une caméra (suppposée fixe) en temps réel, de manière robuste quelles que soient les conditions (éclairage, fond) est un problème difficile non encore résolu. Nous proposons dans cet article une nouvelle approche pour résoudre ce problème, basée sur l'utilisation d'un modèle statistique de suivi. Ce modèle suit le visage à partir de deux informations : la couleur des pixels du visage et la réponse de filtres adaptés répondant prioritairement sur certaines parties d'un visage (yeux, bouche, nez, sourcils). Il est adapté à chaque image en utilisant l'algorithme EM avec un nombre d'itérations fixe, et en utilisant les paramètres passés comme prior. Le suivi temps réel obtenu est indépendant du fond, robuste aux changements d'éclairage, aux occultations partielles du visage, aux changements de pose de la tête.

Mots clefs

Suivi de visage, Temps réel, EM, Modèle Statistique

1 Introduction

Détecter et suivre le visage d'une personne par une caméra (suppposée fixe) en temps réel, de manière robuste quelles que soient les conditions (éclairage, fond) est un problème difficile non encore résolu. Nous proposons dans cet article une nouvelle approche pour résoudre ce problème. La détection est réalisée par un détecteur de de visage à base de Réseaux de Neurones modulaires. Le suivi est basé sur l'utilisation d'un modèle statistique. Il s'agit d'un modèle génératif qui doit expliquer au mieux les observations obtenues par prétraitement. Il est adapté à chaque image en tenant compte du passé et de l'image courante, en utilisant l'algorithme EM.

2 Prétraitements

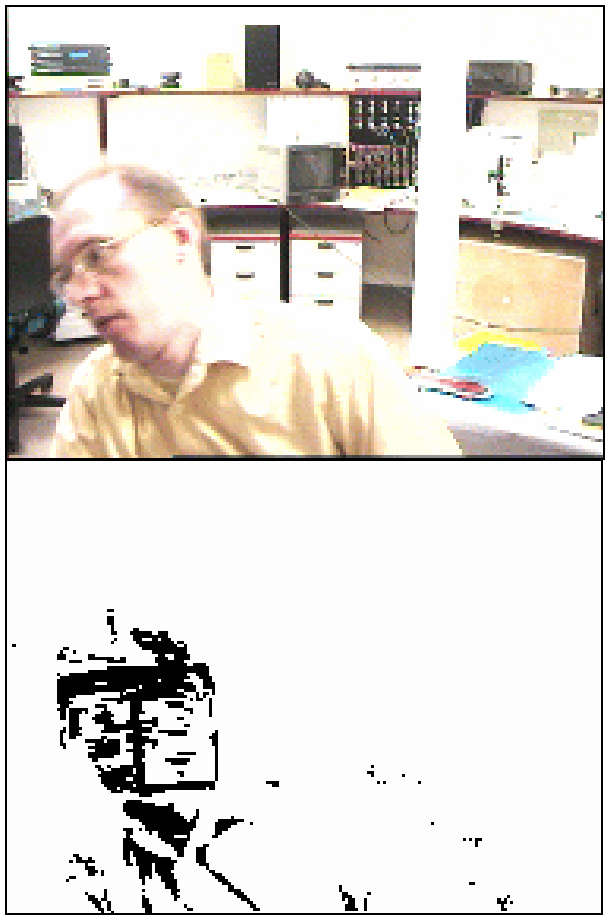

Figure 1 : *Détection de teinte chair et élimination du fond*

Deux prétraitements sont d'abord appliqués sur l'ensemble de l'image acquise. Le premier est l'élimination des pixels qui ne sont pas de teinte chair, par

comparaison avec une table des teintes chair existantes, couvrant toutes les couleurs de peau dans toutes les conditions d'éclairage raisonnables (naturel ou artificiel). Le second est l'élimination des pixels de fond. Pour ce faire, l'image courante est soustraite à une image de référence, acquise à l'initialisation du programme. Cette image de référence est constamment mise à jour sauf dans un rectangle autour du visage lorsque celui-ci est suivi. La figure 1 présente un exemple de résultat de ces prétraitements.

3 Les filtres

Pour pouvoir suivre un visage dans toutes les conditions, la couleur ne suffit pas et des indices de forme sont nécessaires. Nous utilisons des filtres logiques rapides, correspondants à des détecteurs de barres horizontales sombres entourées de zones plus claires (voir figure 2, les zones vertes correspondent aux zones plus claires comparées à la zone rouge) . Ils sont appliqués à l'image à deux échelles différentes, correspondant à une barre respectivement de hauteur 1 ou 2 pixels .

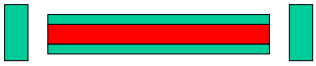

Figure 2 : Filtres détecteurs de barres

Les figures 3 et 4 montrent un exemple de résultat des filtres. Ils répondent prioritairement sur les yeux (et les sourcils) et la bouche, ainsi que sur le nez.

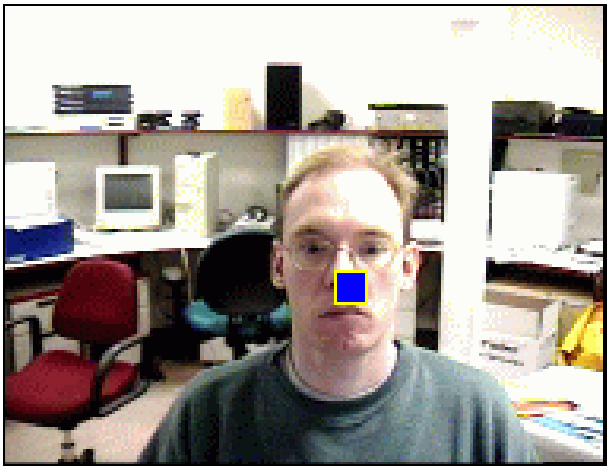

Figure 3 *: image originale*.

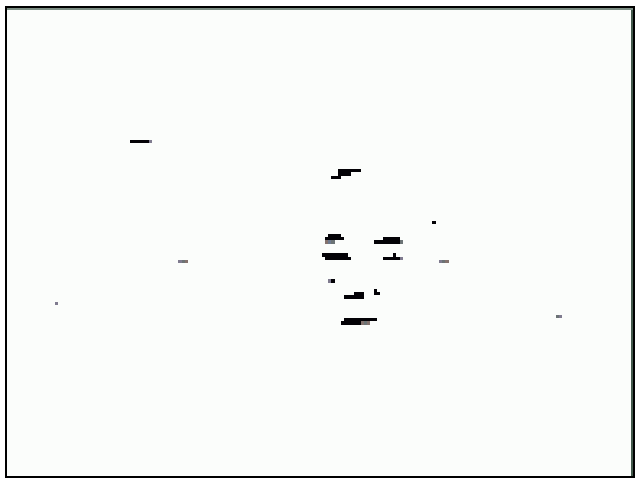

Figure 4 : *résultat du filtrage*

4 Détection de visage

Nous ne détaillerons pas dans cet article le réseau de neurones modulaire détecteur de visage qui a déjà été décrit dans [1]. Ce détecteur a de très bonnes performances (taux de détection de 81% pour une seule fausse alarme sur les 130 images de la base de test standard du CMU numéro 1), mais reste néanmoins trop lent pour une utilisation complète sur toute l'image. C'est pourquoi l'image est découpée en seizes régions égales. A chaque image, une région contenant un nombre suffisant de pixels de teinte chair et de pixels qui ne sont pas du fond est choisie. L'algorithme assure également que si plusieurs zones répondent au critère, elles sont choisies les unes après les autres. Le détecteur de visage est alors uniquement utilisé dans la zone sélectionnée. La détection est ainsi réalisée à 25 image par seconde. Le visage détecté est utilisé pour initialiser le modèle de suivi.

5 Le modèle statistique de suivi

Le modèle est un modèle génératif qui doit expliquer les observations, c'est-à-dire les pixels de teinte chair qui ne sont pas du fond, représentés par leur position (x, y) et leur chrominance (u,v dans l'espace de couleur YUV). Il y a donc un vecteur, noté X, par pixel de teinte chair (leur nombre est noté I). Il doit également expliquer les pixels détectés par les filtres (leur nombre est noté J). Ceux-ci ne sont représentés que par leur vecteur position noté Y :

$$X_i = \begin{bmatrix} x_i \\ y_i \\ u_i \\ v_i \end{bmatrix} \qquad Y_j = \begin{bmatrix} x_j \\ y_j \end{bmatrix}$$

Le modèle statistique comporte donc deux parties : un modèle de génération des pixels de teinte chair et un modèle de génération des pixels détectés par les filtres. Chaque modèle est formé de deux composantes, une pour le visage et une pour le fond. Le modèle de chair pour le visage est composé d'une gaussienne sur la position, couplée à un histogramme H_c pour la couleur, similaire à celui utilisé dans [2]. Il a pour probabilité :

$$P_c(X) = \frac{1}{2\pi}\lambda \exp\left(-\frac{1}{2}\lambda\left[(x-x_0)^2 + (y-y_0)^2\right]\right) H_c(u,v)$$

où x_0, y_0 est la position du visage et λ un paramètre fixe correspondant à la taille du visage, initialisé par le détecteur de visage. Pour le fond, le modèle est une probabilité faible et constante ε pour la position couplée à un historamme H_f :

$$P_f(X) = \varepsilon H_f(u,v)$$

Le modèle de filtres est formé de deux histogrammes de position par rapport au centre du visage (un histogramme horizontal H_h et un histogramme vertical H_v) :

$$P(Y) = H_h(n(x) - n(x_0))H(m(y) - m(y_0))$$

où les fonctions n et m correspondent à la division de l'image en N cellules horizontales et M cellules verticales. Le modèle de fond pour les filtres est une simple probabilité faible uniforme. Dans le modèle statistique total, les paramètres sont la position et les quatres histogrammes. Ils sont adaptés à chaque image.

Pour tenir compte de l'image précédente, et éviter que l'adaptation des histogrammes ne diverge, nous introduisons un prior dans le modèle. Pour la position il s'agit d'un simple prior gaussien centré sur la précédente position. Un modèle dynamique plus sophistiqué peut être introduit en remplaçant la position précédente par une prédiction de la position courante. Pour les histogrammes nous utilisons comme prior l'exponentielle de la divergence de Kullback-Leibler entre l'histogramme courant et l'histogramme à l'image précédente, ce qui donne pour l'histogramme de couleur du visage :

$$P_{prior}(H_c(u,v)) = \exp\left[\alpha \sum_{u,v} H_c^{précedent}(u,v)\log(H_c(u,v))\right]$$

Le même prior est appliqué pour chacun des trois autres histogrammes .

6 Adaptation du modèle

Le modèle est initialisé à chaque image aux paramètres obtenus à l'image précédente. Le modèle est ensuite adapté aux observations de l'image courante, en utilisant l'algorithme EM [3]. Pour le modèle statistique choisi, les variables cachées correspondent à l'affectation de chaque vecteur X ou Y au visage ou au fond. A chaque image, un nombre fixe d'itérations (dans nos exemples, 10 itérations) de l'EM sont appliquées. Chaque itération est décomposée en deux phases.

La première phase (étape E) est la calcul d'un certain nombre de valeurs moyennes en utilisant les paramètres obtenus à l'itération précédente et les observations. Les probabilités pour chaque vecteur X_i ou Y_j qu'il appartienne au visage sont d'abord estimées. Ces probabilités sont notées z_i et z_j . Ensuite, les moyennes suivantes sont calculées :

$$K_c(u,v) = \frac{1}{I}\sum_i z_i \,\delta(u_i - u)\delta(v_i - v)$$

$$K_f(u,v) = \frac{1}{I}\sum_i (1 - z_i)\delta(u_i - u)\delta(v_i - v)$$

$$K_h(n) = \frac{1}{J}\sum_j z_j \delta(n(x_j) - n)$$

$$K_v(n) = \frac{1}{J}\sum_j z_j \delta(m(y_j) - m)$$

$$x_c = \frac{1}{I}\sum_i z_i x_i \qquad y_c = \frac{1}{I}\sum_i z_i y_i$$

La seconde phase (étape M) est la phase de réestimation des paramètres du modèle par maximisation d'un fonction auxiliaire (dénotée Q). Pour les histogrammes de couleur, la réestimation est simplement la suivante :

$$H_c(u,v) = \frac{K_c(u,v) + \alpha H_c^{précédent}(u,v)}{\beta}$$

où β est une constante de normalisation. La réestimation de H_f est similaire. La réestimation de la position se fait indépendamment en x et en y. Pour trouver la nouvelle position, toutes les positions (en x ou en y) correspondant aux centres des cellules des histogrammes de position sont testées. La position pour laquelle la fonction Q est maximum est choisie. Ensuite l'histogramme horizontal est réestimé par la formule suivante (la réestimation est similaire pour l'histogramme vertical) :

$$H_h(n) = \frac{K_h(n - n(x_0)) + \alpha H_h^{précédent}(n)}{\beta}$$

A chaque image, le nombre de pixels chair ou résultants des filtres estimés comme appartenant au visage est calculé. Si ce nombre est trop faible (indiquant un suivi défaillant), le suivi est interrompu et la détection relancée.

7 Résultats

Les images acquises sont au format QCIF (192x144 pixels), en codage YUV. La détection et le suivi se font en temps réel (25 images par seconde) sur un Pentium 4 à 1.7Ghz. Le suivi est robuste au changements d'éclairage, de fond, à la présence de distracteurs dans l'image (voir figure 5), aux changements de pose du visage (voir figure 6) et aux occultations partielles du visage (voir figure 7). Certaines configurations de mains, auxquelles les filtres sont sensibles, peuvent néanmoins leurrer le suivi.

Figure 5 : *suivi avec autre visage et main dans le champ de la caméra*

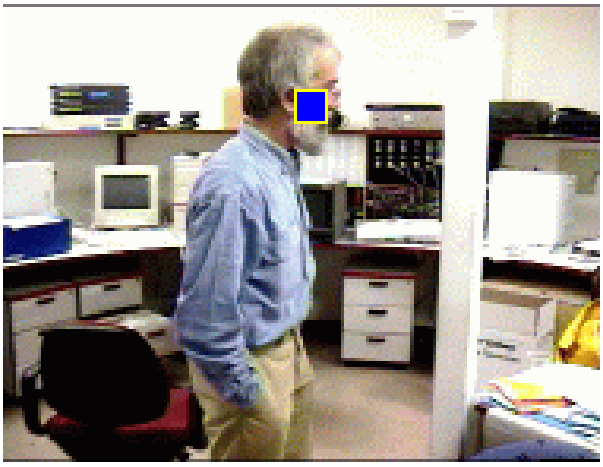

Figure 6 : *suivi pour visage de profil*

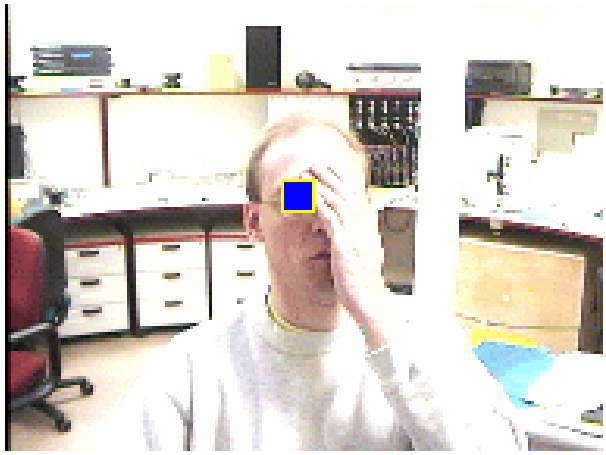

Figure 7 : *suivi avec occultation partielle du visage*

8 Conclusion

Le système est robuste aux occultations partielles, même importantes (occultations par une main par exemple), aux changements brusques d'éclairage, à la présence d'autres visages tant qu'ils n'occultent pas le visage suivi. Il est capable de suivre un visage de taille et orientation variables, tant que la bouche et au moins un œil apparaissent, et ceci quels que soient le fond et les vêtements portés par la personne, en temps réel. Seuls quelques cas de configurations de mains peuvent leurrer l'algorithme. Cela peut être améliorer par une plus grande sélectivité des filtres.

.

Références

[1] Féraud R., Bernier O., Viallet J. E. et Collobert M. "A Fast and Accurate Face Detector based on Neural Networks » dans IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol 23 (2001) 42-53.

[2] Schwerdt K. et Crowley J. L.: "Robust Face Tracking Using Color" dans *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, FG2000, 28-30 march 2000, Grenoble, France.

[3] Dempster A.P., Laird, N.M. et Rubin, D.B.:"Maximum-likelyhood from incomplete data via the EM algorithm". Dans *Journal of the Royal Statistical Society*, Vol 39(1977), 1-38.