LOCALISATION DES PERSONNES DANS UNE SEQUENCE VIDEO

ROJBI Anis SHMITT Jean-Claude GEORGES Alquie BOISSONADE Patrick

& France Telecom R&D DMI/SGV

38,40 rue du Général Leclerc

92794 Issy les Moulineaux Cedex 9

anis.rojbi@rd.francetelecom.com

Résumé

Il est important d'accorder de l'intérêt à la phase d'initialisation de l'algorithme de segmentation par contour actif car de cette étape dépend la qualité finale de la segmentation. En effet, l'étape d'initialisation correspond au calcul de la position de l'objet d'intérêt sans connaissance à priori de la scène, c'est-à-dire sans avoir aucune idée de l'endroit où elle se trouve. Elle peut être approximative dans la mesure où elle peut être raffinée ensuite grâce à l'algorithme de contour actif B-snake . Dans ce contexte, nous proposons une méthode de localisation des personnes dans une séquence vidéo, en temps réel. L'avantage de l'approche proposée réside dans sa simplicité d'implémentation et dans son efficacité pour donner une segmentation grossière ainsi qu'une localisation précise de l'objet d'intérêt. La démarche adoptée consiste en une approche coopérative. Elle exploite la propriété vidéo la plus pertinente (couleur et mouvement) associée à l'information extraite du modèle de forme 2D. Cette méthode a été spécialisée dans la phase d'initialisation dans le cas d'analyse de séquences vidéo provenant des communications visiophoniques (caméra fixe).

Mots clefs

Localisation, teinte chair, couleur, mouvement, Initialisation, B-snake, interpolation, visiophone.

1 Introduction

Cet article décrit un algorithme de segmentation vidéo pour l'extraction des personnes dans un environnement inconnu, tout en respectant les contraintes d'environnement de l'application. L'importance de l'approche proposée réside dans sa simplicité et son efficacité, il est basé sur l'exploitation d'un algorithme de détection de visage associé à la segmentation par le mouvement et à des techniques de traitement (interpolation, modèle 2D, ACP) pour aider la phase décisionnelle (algorithme de B-snake) à converger d'une manière plus efficace, stable et rapide. La démarche adoptée consiste à détecter, tout d'abord, une zone d'intérêt par la technique de teinte chair ou la segmentation de mouvement-couleur suivant un critère de pertinence bien défini. Ensuite, les paramètres les plus importants sont évalués en utilisant une méthode itérative basée sur une analyse par composantes principales. Plusieurs vues sont généralement utilisées pour la mise au point. Les autres paramètres sont ensuite déterminés grâce à l'algorithme de suivi temporel. La seconde étape consiste à comparer la projection du modèle avec les données extraites de l'image 2D. Cette comparaison est faite en se basant sur deux types de primitives : des points d'intérêt (correspondant aux sommets du modèle par exemple) ou des segments (correspondant par exemple aux arêtes) de droites. À chaque primitive est ensuite associée une mesure permettant de savoir si la pose est en accord ou non avec l'image 2D.

Finalement, une méthodologie d'interpolation nous permet de décider du contour approximatif de l'objet (êtres humains) à détecter. La méthode de segmentation B-snake prend en compte la topologie de la segmentation dans la phase de la localisation prédéterminée.

La littérature, porte peu des travaux sur le sujet (à ma connaissance). Toutefois, plusieurs travaux portant sur la détection de visage et de mouvement sont souvent utilisés dans des applications de codage, tracking, indexation, surveillance…

Parmi les méthodes existantes, on cite :

- L'approche le plus répondue est sans doute, celle utilisant un réseau de neurones pour classifier les pixels de l'image, en tant que visage ou non-visage. L'inconvénient de cette approche réside dans le temps de calcul qui ne permet pas souvent de faire des traitements en temps réel.
- L'approche qui consiste à déterminer des points d'intérêts (maxima locaux d'un filtre gaussien aux dérivées secondes), en partant de ces derniers, on réalise une détection de contours qui seront examinés pour être groupés en régions. Le groupement est basé sur leur proximité et leur similarité en orientation et épaisseur. A partir de chaque région, on définit alors un vecteur dont on calculera la moyenne et la matrice de

covariance par rapport aux différents vecteurs modèles. Le critère d'appartenance à un élément du visage s'appuie sur la distance de Mahalanobis, les différents candidats sont alors groupés en se basant sur un modèle de connaissance indiquant leur position relative. Chaque composant du visage est enfin analysé avec un réseau bayesian. L'intérêt de cette méthode est qu'elle peut détecter des visages dans diverses poses.

- Récemment, des études ont montré que la variabilité de la couleur de la peau tenait plus de la différence d'intensité plutôt que de la chromaticité. On peut utiliser plusieurs types d'espaces de couleurs (RGB, RGB normalisé, HSV, YcrCb, YIQ, YES, CIE XYZ, CIE LUV.).

  L'information de couleur est un outil efficace pour identifier des zones du visage cependant, cela n'est pas utilisable lorsque le spectre de couleurs varie de manière significative entre l'arrière plan et le visage. D'autres études ont cependant montré que la combinaison d'analyse de formes, de segmentation de couleurs et d'informations de mouvement pour la localisation et le suivit de visages dans des séquences conduisent à des résultats intéressants.

2 Détection de visage :

L'approche proposée consiste à effectuer tout d'abord une transformation d'espace de couleur RVB→HSV, ensuite d'égaliser l'histogramme de saturation. Cela se traduit par élimination des ombres, ce qui rend la teinte plus pertinente.

Le principe de base de la méthode repose sur une fusion numérique des informations de Teinte et de Luminance suivant une stratégie de pertinence de teinte. Il s'agit donc maintenant de réaliser ce mélange. La démarche consiste à pondérer l'information de Teinte par un coefficient β qui doit être directement lié à la Saturation par une loi exponentielle [4].

$$G_{Couleur} = \beta.G_T + (1-\beta).G_L \qquad \beta = 1 - e^{\lambda s}$$

$$G_{couleur} = G_T - (G_T - G_L) * e^{\lambda s} \qquad \beta \in [0,1] \quad et\ \lambda \le 0$$

G_T et G_L désignent respectivement les gradients Teinte et Luminance.

Le gradient couleur sera essentiellement constitué du gradient Teinte lorsque la Teinte sera pertinente et des gradients Luminance quand, au contraire, la Teinte ne sera pas significative. Le figure 2(a) montre le résultat obtenu par le détecteur de visage, on note bien la présence de quelques trous. En calculant le nombre de ces trous, nous décidons d'appliquer ou non la méthode de clustering [5].

$E = X - N$.

E, est le nombre d'Euler, X, est le vecteur des composants connexes et N, est le nombre de trous.

L'algorithme de classification sur la segmentation obtenue par critère de couleur consiste à :

1. Initialiser *N* centres de gravité à une position arbitraire dans l'espace des attributs ;
2. Affecter les pixels de l'image à la classe dont le centre de gravité est le plus proche du point qui est associé au pixel dans l'espace de représentation ;
3. Mise à jour des centres de gravité, en calculant le centre de gravité des nuages correspondant à chaque classe construit lors de l'étape 2;
4. Si les centres de gravité se sont déplacés lors de l'étape 2, alors l'algorithme boucle à l'étape 2 pour effectuer une nouvelle itération.

Si une classe présente une variance intraclasse supérieure à un seuil fixe, alors elle est divisée en 2. Si les centres de gravité de 2 classes sont distants d'une valeur inférieure à un seuil fixe par l'utilisateur, alors les 2 classes correspondantes sont fusionnées en une seule.

Cette méthode de classification vise en fait à minimiser la variance intraclasse définie par

$$S_C = \sum_{j=1}^{N} \sum_{pi \in Cj} \left\| c(p_i) - \mu_j \right\|^2 .$$

Le figure 1, montre que $\forall$ l'initialisation des centres de gravité, l'algorithme converge vers un minimum local de S_C.

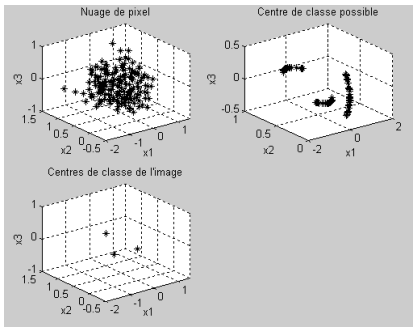


Figure 1– *Principe de la méthode de « clustering »*

L'extraction des paramètres utiles de la région segmentée nécessite :

1. Le calcule du centre de masse $C(\overline{x},\overline{y})$:

$$\overline{x} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} jR[i,j] \qquad \overline{y} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} iR[i,j]$$

$N*M$ est le nombre de pixel de région segmentée et $R[i,j]$ c'est une matrice de dimension $N*M$.

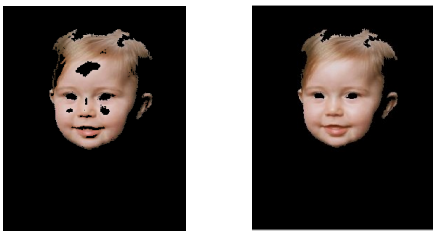
Le calcul de l'angle de rotation θ

$$\theta = B * \tan \frac{2 \sum_{i=1}^{M} \sum_{j=1}^{N} x'_{i,j} \, x'_{i,j} R[i,j]}{\sum_{i=1}^{M} \sum_{j=1}^{N} \left[\left(x'_{i,j} \right)^2 - \left(y'_{i,j} \right)^2 \right] R[i,j]}$$

$$B = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(x_{i,j}' \right)^2 R[i,j]$$

Et $x'_{i,j} = x_{i,j} - \bar{x}$ $y'_{i,j} = y_{i,j} - \bar{y}$

Une fois que nous avons déterminé le centre de masse et l'angle de rotation, nous pouvons redimensionner le modèle de forme et déterminer son inclination par rapport à l'axe principale.



(a) (b)

Figure 2 – *(a) : détection de visage par la méthode de segmentation couleur*
(b) : Traitement de l'image extraite.

Le figure 2(b) montre, l'amélioration de la qualité de détection, grâce aux traitements locaux (d'égalisation d'histogramme et de clustring) appliqués sur l'image segmentée.

3 Segmentation par le mouvement

L'apport de la vidéo dans l'analyse de scène est certainement la possibilité d'utiliser le mouvement comme information supplémentaire à celle de la couleur.
Les techniques de détection de mouvement peuvent être classées de différentes façons :

- celles basées sur des différences d'images.
- celles utilisant un test de vraisemblance.
- celles qui améliorent les décisions ponctuelles par des traitements spatiaux.
- celles qui prennent explicitement en compte les variations d'éclairement.

Dans notre cas, nous proposons d'utiliser l'approche de segmentation de mouvement qui permet de commuter selon le cas, entre l'approche simple de différence d'images et celle de test de Maximum de vraisemblance [3]. Notre choix a été validé pour être un compromis entre les contraintes environnementales (luminance, temps de calcul…) du projet. En effet l'algorithme consiste à appliquer une méthode simple de différence d'image quand la teinte est pertinente, pour raffiner la qualité de détection, alors que dans le cas contraire, on applique la méthode de test de Maximum de vraisemblance avec modèle de luminance supposée constante, pour aider l'algorithme à converger vers une solution acceptable.
La méthode de différence simple d'images consiste à calculer tout d'abord les différences entre l'image courante I_t, l'image précédente I_{t-1} et l'image suivante dans la séquence I_{t+1}. Ensuite un opérateur ET logique est appliqué sur le résultat des deux différences.

$$O_v = (I_t - I_{t-1}) \times (I_t - I_{t-1})$$

O_V est un vecteur qui représente les objets en mouvement.
La méthode de Maximum de vraisemblance *(Maximum Likelihood*) est une méthode similaire à une différence d'image, mais plus robustes, car elle fait intervenir le voisinage du pixel considéré dans la prise de décision. C'est pour cette raison qu'on privilège son exploitation dans le cas où, l'information obtenue par la teinte n'est pas pertinente.
Le principe de la méthode est de maximiser soit l'hypothèse de mouvement, soit l'hypothèse opposée, d'absence de mouvement.
Soit, $I = I(\mu) + \sigma$, le modèle de la luminance [5].
La fonction de vraisemblance s'écrit suivant l'équation

$$\Re_{t-1} = \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right)$$

Par simple dérivation par rapport à μ_m et σ^2 on obtient facilement :

$$\hat{\mu}_m = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Soit F1 et F2 deux fenêtres de taille $N = m*m$, centrées sur le pixel (x_0, y_0) de distribution g1 et g2.
Dans ce cas on a deux hypothèses :

- H_0 : g_1 *et* g_2 proviennent de la même distribution $N(\mu_0, \sigma_0)$ ⇒ Absence de mouvement.
- H_1 : g_1 *et* g_2 proviennent de deux distributions différentes $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$ ⇒ Objets en mouvement.

Le rapport de vraisemblance est défini par :

$$\lambda = \frac{\Re_t(H_1)}{\Re_{t-1}(H_0)} = \frac{\Re_{F1} \times \Re_{F2}}{\Re_{F1 \cup F2}}$$

On pose : $\lambda' = \ln \lambda \quad \Rightarrow \quad \lambda' = \frac{n}{2\sigma} |\mu_1 - \hat{\mu}_2|$

Si ce rapport est supérieur à un seuil, l'hypothèse H_1 est retenue, sinon l'hypothèse H_0 est validée. Pour simplifier le problème de seuil, on propose d'utiliser un intervalle de

confiance [3].Ce qui permet une discrimination très efficace des zones de mouvement dans une image, avec l'avantage de ne pas être dépendant d'un seuil fixe.
On applique cet algorithme sur les 3 premières images de la séquence vidéo. Le suivi temporel est assuré par l'algorithme de contour actif (B-Snake).

4 Extraction des paramètres

La phase initiale de segmentation réalisée par l'algorithme de détection de visage et la méthode de segmentation de mouvement permettent d'extraire les contours d'objets d'intérêts et leurs arêtes. Un modèle de cible est projeté dans le plan de l'image. Ensuite, la projection du modèle est comparée avec les données extraites. Cette comparaison est faite en se basant sur deux types de primitives : les points d'intérêt ou des segments de droites. A chaque primitive est associée une mesure permettant de savoir si la pose est en accord ou non avec le modèle 2D. Pour cela, il faut commencer par trouver où se situent les primitives dans l'image 2D [1]. Pour les points d'intérêt, la mesure correspond alors à la différence entre les coordonnées de projection du point d'intérêt et le point image correspondant. Dans le cas des segments de droite, les segments projetés du modèle sont discrétisés en un ensemble de points. La mesure de comparaison entre l'image 2D et le segment dépend alors de deux distances : celle entre le point d'arrête et le segment de droite et celle de projection orthogonale du point d'arrête et le point modèle. Cette mesure est minimisée par un algorithme de type Gaussien-Newton.
Afin de rendre l'algorithme plus robuste à des changements brusque d'apparence, une base de modèles est créée (indépendamment de l'algorithme principal), celle-ci englobe progressivement tous les aspects que peut prendre la cible. L'utilité d'une telle base est de retrouver la cible lorsque celle-ci a été perdue (si la segmentation initiale n'est pas adéquate).
Pour représenter des objets extraits à l'aide de modèle 2D, La modélisation par contour présente un grand intérêt, puisque si le contour d'un objet est identifiable dans une séquence d'image il sera possible de trouver sa pose. Un contour est plus intéressant qu'une silhouette dans la mesure où il existe une structure sous-jacente. En effet, alors que la silhouette n'est composée que d'un ensemble de points d'arête sans structure, le contour est lui constitué d'un ensemble de points connexes ayant un ou deux voisins. Nous utilisons les courbes B-splines cubique pour approximer les contours d'objets [2]. La représentation se base sur un certain nombre de points de contrôle qui donnent la forme globale de la courbe. Une base de fonctions polynômiales permet de lisser le contour en fonction de ces points. La base de fonctions utilisées est celle des splines cubiques. Elle permet d'approximer la forme de l'objet d'intérêt (l'être humain). La figure 3 illustre le principe. L'un des intérêts principaux des B-splines est qu'elles peuvent être déformées en déplaçant les points de contrôle. Ceci permet de s'adapter à des déformations d'objets dans l'image 2D.

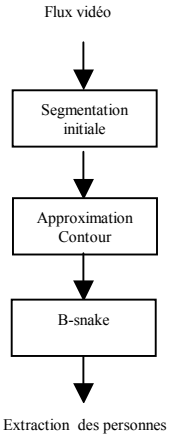


Figure 3 – *Algorithme de segmentation vidéo*

5 Résultats

Le figure 4 illustre les résultats de l'algorithme sur une séquence Akiyo. Une grande partie du processus de traitement est effectué par la méthode de segmentation initiale présentée dans les parties 2 et 3 de l'article. Cela se traduit par un gain important en temps de calcul.

Pour la plupart des séquences, l'algorithme donne un résultat comparable à celle de la séquence Akiyo : une localisation précise et un temps de calcul acceptable.
Pour une cadence de traitement de 4 images par seconde, pour des images QCIF avec un programme non optimisé sur un PC (983 MHz).
L'algorithme ne permet pas de donner des bons résultats quand le niveau de saturation des couleurs est très bas et l'objet d'intérêt est immobile. (Pratiquement, la probabilité de rencontrer ce cas est très faible).

6 Conclusion

Un algorithme de localisation des personnes a été développé, dans le contexte d'applications de visiophone fixe. L'avantage de l'algorithme se manifeste par sa simplicité et son efficacité pour donner des résultats de segmentation optimale pour la plupart des cas de l'application considérée.

Références

[1] D.B.Gennery.Visual Tracking of Known Three-Dimensional Objects, *international Journal of Computer Vision*, Vol. 7, No. 3, pp. 243-270, April 1992.

[2] D. Koller, J. Weber, T. Huang, J. Malik,G.Ogasawara, B. Rao, and S.Russell. Towards Robust Automatic Traffic Scene Analysis in Real-Time, *International Conference on Pattern Recognition*, 12th, Jerusalem, Israel, Vol. I, pp. 126-131 October 9-13, 1994.

[3] Y. Z. Hsu, H. H. Nagel, G. Rekers. New likelohood test methods for change detection in image sequences, *IEEE CVGIP 26*, pages 73-106, 1984.

[4] T.Carron. Segmentation d'images couleur dans la base Teinte Luminance Saturation : approche numérique et symbolique, *Thèse de l'Université de Savoie* soutenue en décembre 1995.

[5] Z.Wu and R.Lcahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11): 1101--1113, October1993.

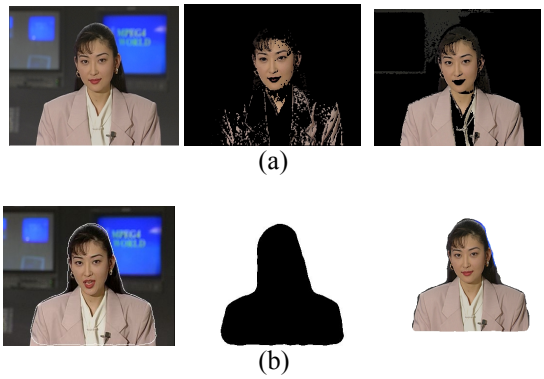


Figure 4 – *(a) : les résultat obtenue par la segmentation initile de couleuret de Mouvement. (b) :convergence rapide du B-snake en exploitant les résultats de segmentation intiale*