Détection de personnes dans les vidéos pour leur immersion dans un espace virtuel

Pascal BERTOLINO, Guillaume FORET, Denis PELLERIN

Laboratoire des Images et des Signaux
ENSIEG, Domaine universitaire, BP 46
38402 Saint Martin d'Hères cedex, France
Pascal.Bertolino, Guillaume.Foret, Denis.Pellerin@lis.inpg.fr

Résumé – Ce papier présente une application permettant d'extraire des personnes évoluant dans des séquences filmées en intérieur ou en extérieur par une caméra fixe. Ces personnes sont incrustées en temps réel dans un milieu virtuel de bande dessinée où elles peuvent interagir en tant qu'acteurs et spectateurs. L'originalité de notre méthode repose sur deux principaux points. Tout d'abord la détection de changement est effectuée en combinant deux types de masques : un masque région et un masque contour. Ensuite ces masques sont obtenus à l'aide d'une image de référence, qui est construite puis mise à jour régulièrement. Cette dernière prend ainsi en compte les différents changements qui apparaissent souvent lorsque la séquence dure plusieurs heures. Les résultats montrent que les personnes sont correctement et rapidement extraites.

Abstract – This paper presents an application developed to extract people in indoor or outdoor environments using a fix camera. People are then incrusted in real time in a dynamic comics environment in which they will interact as actors or/and witnesses. The originality of our method relies on two main points. First, the change detection method combines two kinds of masks: region and contour masks. Then, in order to detect people who are moving or not, a reference image is introduced. It is built and updated, taking into account changes that often occur when the sequence lasts several hours. Results show that people are well and quickly extracted.

1 Introduction

Le travail présenté s'inscrit dans le cadre d'un projet européen de l'IST [1] : Art-live [2](IST 10942). L'objectif de ce projet est d'incruster dans un environnement de bande dessinée des personnes filmées « dans la rue » en temps réel, et de les faire interagir avec l'environnement de la BD, selon un scénario préconçu. Cette application doit pouvoir fonctionner avec le minimum de contrôle pendant toute une journée. La qualité de l'incrustation dépend en grande partie de l'extraction en temps réel et de la qualité des masques des personnes qui passent ou s'arrêtent dans le champ de la caméra qui est fixe.

Ce papier traite de l'aspect segmentation. Aucun dispositif spécial (blue screen dans le fond, capteurs ou marques sur les personnages) ne permet de réaliser la segmentation dans des conditions optimales. L'une des principales contraintes imposées par ce projet est le respect du temps réel (au minimum 8 images 352x288 pixels par seconde). Les traitements utilisés doivent donc être simples et efficaces.

La première partie (section 2) de ce document correspond à la construction du masque représentant la personne. Celui-ci est obtenu par combinaison de deux opérateurs pour être moins sensible à la présence d'ombres dans la scène. La seconde partie (section 3) présente la gestion de l'image de référence, qui permet d'extraire toute personne mobile ou immobile dans la séquence vidéo. Finalement des résultats sont présentés et commentés.

[1] Information Society Technologies

[2] Architecture and authoring tools prototype for Living Images and Video Experiment

2 Construction des masques

La caméra étant fixe, une solution simple consiste à utiliser une image représentant la scène en l'absence de tout individu. L'utilisation de cette image, communément appelée image de référence [1, 2], rend immédiate la détection de présence d'une personne. Nous considérons dans un premier temps que cette image de référence est disponible, nous présenterons au cours de la partie suivante la manière dont cette image est obtenue.

2.1 Combinaison d'un masque région et contour

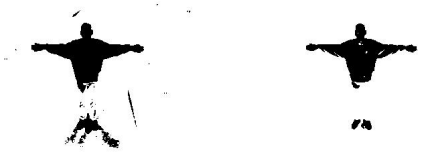
Une approche commune [3] consiste à calculer la différence D entre l'image courante I et l'image de référence I_{ref} pixel par pixel. Cette image différence D est alors seuillée pour former un masque.

D'autres approches [4] effectuent le même calcul à partir de l'image gradient I' de l'image courante et de l'image gradient I'_{ref} de l'image de référence. Plus récemment, dans [5], les auteurs utilisent conjointement l'information couleur et contour.

Afin d'être peu sensible à la présence d'ombre, nous combinons l'information de niveau de gris et l'information contour : pour chaque image de la séquence, deux masques M_1 et M_2 sont calculés, ils sont alors combinés par un *OU* logique pour fournir un seul masque M. Pour chaque pixel de coordonnées (x, y) à l'instant t, nous avons :

$D_1(x, y, t) = |I(x, y, t) - I_{ref}(x, y, t)|$
$D_2(x, y, t) = |I'(x, y, t) - I'_{ref}(x, y, t)|$
si $D_1(x, y, t) \geq \lambda_1$ alors $M_1(x, y, t) = 1$ (masque)
sinon $M_1(x, y, t) = 0$ (fond)
si $D_2(x, y, t) \geq \lambda_2$ alors $M_2(x, y, t) = 1$ (masque)
sinon $M_2(x, y, t) = 0$ (fond)

a) image originale



b) masque obtenu avec un seuil faible

c) masque obtenu avec un seuil supérieur

FIG. 1 – Problème du seuillage avec le masque région

a) masque région

b) masque contour

c) combinaison des masques

d) résultat obtenu après le post-traitement

FIG. 2 – Combinaison du masque région et du masque contour

λ_1 et λ_2 sont deux seuils de décision. Leur valeur est comprise dans l'intervalle $[0, 255]$ puisque nous traitons des images en niveau de gris. Pour calculer les images gradient I' et I'_{ref}, les opérateurs de Prewitt, à la fois rapides et robustes face au bruit, sont utilisés.

Le masque région M_1 est assez sensible à la présence d'ombres dans la séquence. En effet, si on utilise un seuil λ_1 trop faible, l'ombre de la personne apparaît dans le masque (fig. 1.b et 1.c). Ces masques montrent également que les zones peu contrastées ne sont pas extraites correctement. La construction du masque contour M_2 se montre moins sensible aux ombres car celle-ci n'est pas détectée lors du calcul du gradient. La combinaison du masque M_1 (fig. 2.a) avec le masque contour M_2 (fig. 2.b) permet de renforcer le contenu du masque de la personne tout en restant insensible à l'ombre (fig. 2.c). Les seuils suivants ont été utilisés : $\lambda_1 = 20$, $\lambda_2 = 5$.

2.2 Pré et post-traitement

Pour limiter l'influence du bruit d'acquisition, toutes les images traitées (I, I', I_{ref} et I'_{ref}) sont pré-traitées avec un filtre moyenneur 3×3.

Comme le montre la figure 2.c, le masque obtenu par combinaison possède des trous. C'est pourquoi nous appliquons à ce masque un post-traitement en trois étapes. La première relie les pixels extraits suivant la verticale sous certaines conditions (distance, niveau gris). Ce remplissage conditionnel permet de compléter le corps des personnes qui est le plus souvent orienté verticalement [6]. Les trous de petite taille sont alors supprimés à l'aide d'une fermeture morphologique (fig. 2.d). Enfin un étiquetage en composantes connexes permet de supprimer des petites régions parasites.

3 Gestion de l'image de référence

La conception de l'image de référence peut se résumer à l'acquisition de la scène sans objet. Cependant la stabilité de l'illumination ne peut être garantie au fil de la séquence, spécialement pour les scènes extérieures. Cette image doit donc être remise à jour régulièrement. Il est également possible que cette image de référence ne soit pas disponible à l'initialisation, il faut alors pouvoir la construire.

3.1 Principe de notre approche

De la même manière que [7, 8], notre approche est basée sur l'utilisation de deux modes : le premier construit la référence, le second la met à jour. Ces deux modes sont présentés sur le diagramme d'état de la figure 3.

A l'initialisation de l'algorithme, nous décidons s'il faut construire ou seulement mettre à jour l'image de référence. En ce qui concerne la construction, elle s'effectue à partir de la première image de la séquence sur un nombre fixé d'images. Ensuite l'application bascule automatiquement vers le mode « mise à jour ». L'application reste dans ce mode tant que l'image de référence ne subit pas de forte dégradation. Si tel est le cas, le superviseur peut forcer l'application à retourner en mode « construction » afin de réinitialiser la référence.

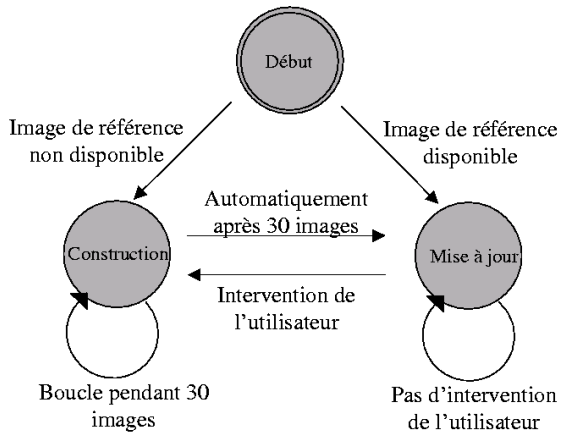


FIG. 3 – Les deux modes de gestion de l'image de référence

3.2 Construction de l'image de référence

Pour construire une image de référence, l'application apprend le contenu de l'image au niveau des zones où la valeur des pixels ne varie pas au cours du temps [9]. La valeur de ces pixels enrichit l'image de référence grâce à une somme pondérée entre l'image courante et l'image de référence existante, pour chaque pixel (équation 1) :

$$I_{ref}(p, t+1) = \alpha_p . I(p,t) + (1 - \alpha_p) . I_{ref}(p,t) \quad (1)$$

- $I_{ref}(p,t)$ et $I(p,t)$ sont les valeurs d'intensité du pixel p à l'instant t dans l'image de référence et dans l'image courante.
- α_p est le coefficient d'adaptation. Si p appartient au fond $\alpha_p \in]0, 1]$, sinon $\alpha_p = 0$.

La valeur d'α_p fixe la vitesse d'apprentissage. Pour apprendre progressivement l'image de référence, nous avons choisi une valeur faible : $\alpha_p = 0.1$. Ainsi seuls les éléments stables et durables sont assimilés dans la référence.

Pour réaliser le calcul de l'équation 1, nous devons connaître les pixels qui appartiennent au fond. Nous utilisons pour cela une carte de stabilité qui est obtenue en calculant la différence entre trois images successives de la séquence ($I(t-1)$, $I(t)$ et $I(t+1)$). Cette carte de stabilité est une image binaire qui distingue les pixels $mobiles$ et les pixels $fixes$ dans ces trois images. Tous les pixels déclarés $fixes$ appartiennent au fond, et l'image de référence est mise à jour uniquement pour ces pixels.

Ce mode d'apprentissage a un inconvénient. En effet, si une personne s'immobilise dans le champ de la caméra, elle est progressivement introduite dans la référence. Nous perdons alors le masque de cette personne. C'est pourquoi nous avons choisi de limiter dans le temps cette phase de construction. Le nombre d'images utilisées pour la construction dépend de la vitesse des entités présentes dans la séquence. Dans le cadre de notre application, nous avons utilisé 30 images successives.

3.3 Mise à jour de l'image de référence

Lorsque la phase de construction est terminée, nous considérons que l'image de référence est fiable. Le mode de mise à jour est alors utilisé pour conserver cette qualité. Au cours de cette mise à jour, les changements locaux, qui sont dus aux faibles modifications dans le fond, et les changements globaux, qui affectent l'image dans son ensemble, sont traités différemment.

3.3.1 Changements locaux

La technique de mise à jour présentée dans cette partie permet de prendre en compte les faibles variations locales dans le contenu de l'image. La contribution de l'image courante au cours de cette mise à jour est également régulée par l'équation 1. La différence avec le mode précédent repose sur le fait que nous n'utilisons pas de carte de stabilité pour déterminer les pixels appartenant au fond. En effet, la qualité de l'image de référence permet d'obtenir un masque correct des entités présentent dans la séquence. L'ensemble de ces entités correspond au premier plan, nous connaissons donc par défaut les pixels du fond.

3.3.2 Changements globaux d'illumination lents ou rapides

Les changements globaux d'illumination arrivent fréquemment dans les séquences intérieures et extérieures. Ces changements affectent dans tous les cas le contenu de l'image de référence. Ces changements sont donc détectés afin d'apporter une correction à la référence.

Un changement rapide d'illumination peut être détecté entre deux images successives. Pour cela une différence globale est calculée pour les images t et $t+1$:

$$\Delta_1 = \frac{\sum_p I(p, t-1) - \sum_p I(p,t)}{N} \quad (2)$$

N correspond au nombre de pixels dans l'image. Si $|\Delta_1| > C_1$ ($C1$ étant un seuil) un changement rapide d'illumination est détecté. La différence moyenne Δ_1 est alors ajoutée à chaque pixel de l'image de référence.

Un changement progressif (lent) d'illumination peut être absorbé par l'équation 1 tant qu'il n'est pas trop important par rapport au coefficient d'adaptation α_p. Si la valeur d'α_p ne permet pas de compenser assez rapidement cette variation dans la référence, la valeur moyenne d'illumination dans l'image de référence devient de plus en plus différente de celle de l'image courante. Ce type de variation lente ne peut pas être détectée entre deux images successives, mais elle peut devenir détectable entre l'image courante et la référence. Ainsi le même principe de détection et de correction que précédemment est utilisé entre l'image courante et la référence, avec un seuil $C2$:

$$\Delta_2 = \frac{\sum_p I(p,t) - \sum_p I_{ref}(p,t)}{N} \quad (3)$$

Si $|\Delta_2| > C_2$ un changement d'illumination lent est détecté. La différence moyenne Δ_2 est alors ajoutée à chaque pixel de la référence.

4 Résultats

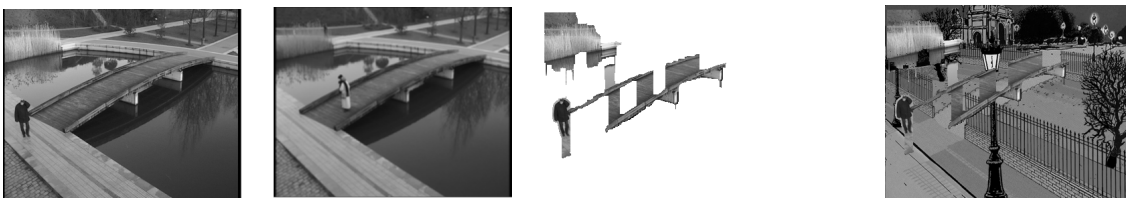
Cette application a été développée en langage C, elle est utilisable sous environnement Unix, Linux et Windows. Elle peut traiter une séquence vidéo à la fréquence de 8 images (352×288) par seconde sur un Pentium III $800MHz$. Les résultats présentés ont été obtenus à partir d'une séquence du projet Art-live (filmée en extérieur).

La planche de résultats présente à trois instants donnés : l'image originale de la séquence, l'image de référence, le masque extrait et un exemple d'incrustation.

La figure 4 représente l'initialisation du traitement. Pour tester la robustesse de la technique de construction de l'image de référence, nous avons choisi aléatoirement une image appartenant à la séquence pour jouer le rôle de la référence initiale (fig 4.b). Cette image contient une personne et son illumination est différente de celle de la première image traitée (fig 4.a). C'est pourquoi le masque obtenu est de mauvaise qualité (fig 4.c).

La figure 5 montrent que plusieurs images sont nécessaires pour apprendre la totalité de la référence. Ce but est atteint après environ une vingtaine d'images (fig 5.b). La qualité de la référence permet alors d'avoir un bon masque (fig 5.c).

La figure 6 montrent que les variations d'illumination sont prises en compte au cours de la mise à jour de la référence. (Dans l'image fig 6.a, le temps s'assombrit)



a) image originale b) image de référence c) masque extrait d) incrustation (© Casterman-Tardi)

FIG. 4 – Image 1 : Initialisation. L'image de référence doit être reconstruite pour améliorer la qualité des masques.

a) image originale b) image de référence c) masque extrait d) incrustation (© Casterman-Tardi)

FIG. 5 – Image 25 : Construction de l'image de référence. Plusieurs images sont nécessaires pour apprendre l'image de référence.

a) image originale b) image de référence c) masque extrait d) incrustation (© Casterman-Tardi)

FIG. 6 – Image 1000 : La variation d'illumination au cours de la journée est prise en compte dans la référence.

5 Conclusion

Nous avons présenté une application qui permet d'extraire en temps réel les personnes présentes dans une séquence vidéo. Cette application utilise une double extraction de masque pour être moins sensible à la présence d'ombres. La technique de remise à jour de l'image de référence, qui lui est associée, lui permet d'être utilisée sur de longues séquences vidéo. Dans la suite du projet, nous travaillerons sur la gestion de l'image de référence afin de limiter l'intervention du superviseur.

Références

[1] G. W. Donohoe, D. R. Hush, N. Ahmed, *"Change detection for target detection and classification in video sequences"* , In Proc ICASSP, 1084–1087, New York, USA, 1988.

[2] P.L. Rosin,T. Ellis, *"Image difference threshold strategies and shadow detection"'*, In 6th British Machine Vision Conference, pages 347–356, Birmingham, England, 1995.

[3] O.S. Wenstop, *"Motion detection for image information"*, Proceedings of 3^{rd} Scandinavian Conference on Image Analysis, p. 381-386, Tromso, Norway, July 1983.

[4] P. Vannoorenberghe, C. Motamed, J-M. Blosseville, J-G Postaire, *"Motion detection for non-rigid objects. Application to pedestrians monitoring in urban environment"*, IEEE International Conference on IMACS, CESA'96, Lille, France, July 9-12 1996.

[5] S. Jabri, Z. Duric, H. Wechsler, A. Rosenfeld, *"Detection and location of people in video images using adaptive fusion of color and edge information"*, Proceedings of 15^{th} International Conference on Pattern Recognition, Volume 4, p. 627-630, Barcelona, Spain, September 2000.

[6] C. Kim, J-N. Hwang, *"A fast and robust moving object segmentation in video sequences"*, IEEE International Conference on Image Processing, Kobe, Japan, October 1999.

[7] I. Haritaoglu, D. Harwood, L. Davis, *"W4 : Real-time surveillance of people and their activities"*, IEEE Transactions on PAMI, vol. 22, no 8, p. 809-830, August 2000.

[8] E. Bruno, *"Détection robuste du mouvement dans des séquences d'images rapides"*, DEA réalisé au laboratoire LSIIT, Université Louis Pasteur, Strasbourg, France, juillet 1997.

[9] N. Hoose, L.G. Willumsen, *"Automatically extracting traffic data from video-tape using the CLIP4 parallel image processor"*, Pattern Recognition Letters, vol.6, no 3, p. 199-213, August 1987.