Extraction rapide et robuste d'objets vidéo combinant des différences d'images et une image de référence réactualisée

Laurent Bonnaud Alice Caplier Jean-Marc Chassery
Laboratoire des images et des signaux (LIS)
46 avenue Félix Viallet, 38031 Grenoble Cedex, France
<Prénom.Nom>@inpg.fr

Résumé

Cet article décrit un algorithme d'extraction d'objets vidéo développé dans le contexte du projet européen `art.live`[1] *où des contraintes de qualité de segmentation et de cadence de traitement (au moins 10 images/seconde) sont requises. Afin d'obtenir une segmentation de qualité (avec des frontières précises et avec une certaine stabilité temporelle), le processus de segmentation utilise une modélisation par champs de Markov qui prend en compte à la fois des différences d'images consécutives et une image de référence d'une façon unifiée. Les changements temporels de luminance sont prédominants lorsque l'image de référence n'est pas encore (ou pas complètement) disponible, alors que l'image de référence domine pour les objets mobiles faiblement texturés ou pour les objets dont le mouvement cesse.*

La rapidité de traitement de cet algorithme vient du remplacement de certaines itérations markoviennes par des opérateurs morphologiques, qui ne dégrade pas la qualité de la segmentation. Des simulations montrent l'efficacité de la méthode proposée à la fois en termes de qualité et de complexité ($\simeq$ 6 images par seconde pour des images de 352×288 *pixels) sur un processeur d'entrée de gamme.*

Mots clé

traitement de séquences d'images, segmentation vidéo, champs de Markov, image de référence, algorithme rapide

1 Introduction : le projet `art.live`

Le projet européen `art.live` a pour but de développer et démontrer des **espaces narratifs** de réalité mixte (réalité augmentée et réalité virtuelle) situés dans des univers de type bande dessinée. Des artistes et des utilisateurs du système peuvent interagir, grâce à Internet ou plus généralement, tout réseau IP.

Une illustration de ce principe est montrée sur la figure 1 où les personnes et les objets entrant dans le champ de la caméra (figure 1a) sont placés dans un environnement virtuel (figure 1b) visualisé sur un grand écran ou à travers Internet. Ces personnes, filmées en intérieur ou en extérieur, peuvent interagir avec le scénario proposé grâce aux mouvements de leur corps. Elles peuvent aussi interagir avec d'autres personnes situées derrière un écran d'ordinateur.

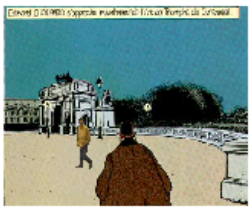
(a)  (b) 

FIG. 1 – (a) : extraction d'objets vidéo. (b) : objets vidéo réels placés dans un environnement virtuel. Image copyright Casterman, J. Tardy et le projet `art.live`.

Dans ce résumé, nous décrivons uniquement la segmentation d'objets vidéo que nous pouvons limiter à un fond fixe, vu les contraintes du projet. De nombreux travaux existent sur ce sujet qui est un problème difficile [6]. Dans la littérature, deux approches principales existent :

- prendre en compte uniquement des différences inter-images [1]. Un défaut majeur de ce choix est qu'aucun changement temporel significatif n'intervient dans la zone de recouvrement d'un objet en mouvement si celui-ci n'est pas assez texturé. Un autre inconvénient est qu'un objet cesse d'être détecté s'il cesse de bouger.
- prendre en compte uniquement la différence avec une image de référence du fond fixe [4]. Ce choix permet de détecter l'objet complet, même s'il est peu texturé ou devient immobile. Le problème qui demeure est la construction et la mise à jour de l'image de référence qui nécessite de nombreuses images et introduit un délai important avant l'obtention d'une segmentation correcte.

Dans notre travail, la segmentation d'objets mobiles est une extension de l'algorithme de détection itératif basé sur les champs de Markov développé dans [3]. Les

[1] `art.live` : IST project 10942, ARchitecture and authoring Tools prototypes for Living Images and Video Experiments

principales modifications portent sur :
- le modèle markovien qui prend en compte 2 observations : la différence entre l'image courante et l'image précédente ainsi que la différence avec une image de référence réactualisée. Le cadre markovien est une façon efficace de prendre en compte différentes sources d'information en vue de prendre une décision. Cela se traduit par la minimisation d'une fonction d'énergie composée de 3 termes d'énergie (au lieu de 2 précédemment). Les changements temporels de luminance sont déterminants lorsque l'image de référence n'est pas encore (ou pas complètement) disponible, alors que l'image de référence prédomine pour les objets mobiles faiblement texturés ou pour les objets qui s'arrêtent momentanément.
- l'accélération de la cadence de traitement : elle vient du remplacement, sans dégradation de la qualité de la segmentation, d'une itération markovienne sur deux (algorithme ICM) par des opérateurs morphologiques (ouverture et fermeture).

La partie 2 décrit l'algorithme de segmentation. La méthode de construction et de mise à jour de l'image de référence est décrite dans la partie 3. La partie 4 présente des exemples d'objets vidéo extraits qui montrent que l'algorithme proposé marche correctement dans la plupart des situations contrôlée utiles pour le projet `art.live`.

2 Extraction d'objets vidéo

2.1 Étiquettes et observations

La détection de mouvement est un problème d'étiquetage binaire dont le but est d'attribuer à chaque pixel ou «site» $s = (x, y)$ de l'image I à l'instant t l'une des deux étiquettes possibles :

$$e(x, y, t) = e(s, t) = \begin{cases} obj \text{ si } s \text{ appartient à un objet} \\ bg \text{ si } s \text{ appartient au fond} \end{cases}$$

$e = \{e(s,t), s \in I\}$ représente une réalisation particulière (à l'instant t) du champ d'étiquettes E. De plus, nous définissons $\Omega = \{e\}$ comme l'ensemble des réalisations possibles du champ E.
Grâce aux hypothèses d'illumination quasi-constante et de caméra fixe, l'information de mouvement est reliée directement aux changements temporels de la fonction intensité $I(s,t)$ et aux changements entre l'image courante $I(s,t)$ et une image de référence $I_{\text{REF}}(s,t)$ qui représente le fond fixe sans aucun objet mobile. Par conséquent, nous définissons 2 observations :

1. l'observation O_{FD} définie comme la différence de 2 images consécutives :
$$o_{\text{FD}}(s,t) = |I(s,t) - I(s,t-1)|$$
2. l'observation O_{REF} définie comme la différence entre l'image courante et l'image de référence :
$$o_{\text{REF}}(s,t) = |I(s,t) - I_{\text{REF}}(s,t)|$$

$o_{\text{FD}} = \{o_{\text{FD}}(s,t), s \in I\}$ et $o_{\text{REF}} = \{o_{\text{REF}}(s,t), s \in I\}$ représentent une réalisation particulière (à l'instant t) des champs d'observation respectifs O_{FD} et O_{REF}.
Pour trouver la configuration la plus probable du champ E étant donnés les champs O_{FD} et O_{REF}, nous utilisons le critère MAP et cherchons $e \in \Omega$ tel que ($Pr[\,]$ étant la probabilité) :

$$Pr[E = e / O_{\text{FD}} = o_{\text{FD}}, O_{\text{REF}} = o_{\text{REF}}] \text{ maximum}$$

ce qui, en utilisant le théorème de Bayes, est équivalent à chercher $e \in \Omega$ tel que :

$$Pr[E = e]\ Pr[O_{\text{FD}} = o_{\text{FD}}, O_{\text{REF}} = o_{\text{REF}} / E = e] \text{ max.}$$

2.2 Fonction d'énergie

La maximisation de cette probabilité est équivalente à la minimisation d'une fonction d'énergie U qui est la somme pondérée de plusieurs termes [5] :

$$\begin{aligned} U(e, o_{\text{FD}}, o_{\text{REF}}) = U_m(e) + \lambda_{\text{FD}} U_a(o_{\text{FD}}, e) \\ + \lambda_{\text{REF}} U_a(o_{\text{REF}}, e) \end{aligned} \quad (1)$$

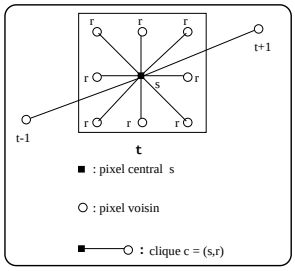


FIG. 2 – Voisinage spatio-temporel et cliques.

Le terme d'énergie du modèle $U_m(e)$ peut être interprété comme un terme de régularisation qui assure l'homogénéité spatiale du masque des objets mobiles et élimine les pixels isolés dus au bruit. Son expression qui résulte de l'équivalence entre champs de Markov et la distribution de Gibbs est :

$$U_m(e) = \sum_{c \in C} V_c(e_s, e_r)$$

où c est une clique quelconque définie sur le voisinage spatio-temporel de la figure 2. Une clique $c = (s,r)$ est une paire quelconque de sites distincts telle que s est le pixel courant et r un quelconque de ses voisins. C est l'ensemble de toutes les cliques. $V_c(e_s, e_r)$ est une fonction potentiel élémentaire associée à chaque clique $c = (s,r)$. Elle prend les valeurs suivantes :

$$V_c(e_s, e_r) = \begin{cases} -\beta_r & \text{si } e_s = e_r \\ +\beta_r & \text{si } e_s \neq e_r \end{cases}$$

où le paramètre positif β_r dépend de la nature de la clique : $\beta_r = 20, \beta_r = 5, \beta_r = 50$ respectivement pour

une clique spatiale, pour une clique temporelle vers le passé et pour une clique temporelle vers le futur. Ces valeurs ont été fixées expérimentalement une fois pour toutes.
Le lien entre les étiquettes et chaque observation (notée de façon générique O) est défini par les équations suivantes :

$$o(s,t) = \Psi(e(s,t)) + n(s)$$

$$\text{où } \Psi(e(s,t)) = \begin{cases} 0 & \text{si } e(s,t) = bg \\ \alpha > 0 & \text{si } e(s,t) = obj \end{cases}$$

et $n(s)$ est un bruit blanc gaussien de moyenne nulle et de variance σ^2. σ^2 est estimée de façon grossière comme la variance de chaque champ d'observation, qui est calculée pour chaque image de la séquence, si bien qu'il ne s'agit pas d'un paramètre arbitraire.
$\Psi(e(s,t))$ modélise chaque observation de sorte que n représente le bruit d'adéquation :
- si le pixel s appartient au fond fixe, aucun changement temporel (autre que le bruit) n'intervient ni dans I, ni dans sa différence avec l'image de référence, si bien que chaque observation est quasiment nulle
- si le pixel s appartient à un objet mobile, un changement intervient dans les 2 observations et chaque observation est supposée proche d'une valeur positive α_{FD} et α_{REF} qui représente la valeur moyenne prise par chaque observation.

Les énergies d'adéquation $U_a(o_{\text{FD}}/e)$ et $U_a(o_{\text{REF}}/e)$ sont calculées selon les relations suivantes :

$$U_a(o_{\text{FD}}, e) = \frac{1}{2\sigma^2_{\text{FD}}} \sum_{s \in I} [o_{\text{FD}}(s,t) - \Psi(e(s,t))]^2$$

$$U_a(o_{\text{REF}}, e) = \frac{1}{2\sigma^2_{\text{REF}}} \sum_{s \in I} [o_{\text{REF}}(s,t) - \Psi(e(s,t))]^2$$

Deux coefficients de pondération λ_{FD} et λ_{REF} sont introduits car le bon comportement de l'algorithme résulte d'un compromis entre tous les termes d'énergie. La valeur $\lambda_{\text{FD}} = 1$ est fixée une fois pour toutes et ne dépend pas de la séquence traitée. La valeur de λ_{REF} est fixée selon la règle suivante :
- $\lambda_{\text{REF}} = 0$ si $I_{\text{REF}}(s,t)$ n'existe pas : quand l'image de référence n'est pas encore disponible au pixel s, $o_{\text{REF}}(s,t)$ n'influence pas le processus de relaxation ;
- $\lambda_{\text{REF}} = 25$ si $I_{\text{REF}}(s,t)$ existe. Cette valeur est élevée car une grande confiance peut être accordée à l'image de référence quand elle existe.

2.3 Relaxation

L'algorithme de relaxation déterministe ICM (Iterated Conditional Modes) [2] est utilisé pour trouver un minimum local de la fonction d'énergie donnée par l'équation (1). Ayant constaté que la diminution la plus importante de la fonction d'énergie se produit dans les quelques premières itérations, nous décidons de n'effectuer que 4 itérations ICM. De plus, entre 2 itérations ICM, une fermeture et une ouverture morphologique sont effectuées sur le champ d'étiquettes E. Il en résulte une augmentation de la cadence de traitement sans perte de qualité puisque les itérations markoviennes restantes continuent à fonctionner directement sur les observations (différences entre images) et non pas sur les champs d'observation binarisés.

2.4 Initialisation

Cet algorithme étant itératif, une initialisation du champ d'étiquettes E est nécessaire. Elle résulte d'un OU logique entre les 2 champs d'observations O_{FD} et O_{REF} binarisés. Cela nécessite 2 seuils de binarisation qui sont choisis en fonction du type de séquence et du système d'acquisition vidéo.

3 Création et mise à jour de l'image de référence

Dans le projet `art.live`, l'environnement réel peut aussi bien être en intérieur ou en extérieur. Dans le cas d'une séquence acquise en intérieur (typiquement un stand de conférence), l'image de référence peut être initialisée lorsque personne ne se trouve dans le champ de la caméra. Par contre, dans le cas d'une séquence acquise en extérieur dans un lieu très fréquenté, il peut s'avérer très difficile de n'avoir personne dans le champ de la caméra. Il devient donc nécessaire de construire l'image de référence peu à peu.
L'image de référence est donc construite image après image, en réutilisant le résultat de la détection décrite précédemment, selon l'équation suivante (s désigne un pixel de l'image et t le temps) :

$$I_{\text{REF}}(s,t+1) = \gamma(s,t) I_{\text{REF}}(s,t) + [1 - \gamma(s,t)] I(s,t+1)$$

$$\text{où } \gamma(s,t) = \begin{cases} 0 & \text{si le pixel est statique et } I_{\text{REF}}(s,t) \text{ n'existe pas} \\ 0.5 & \text{si le pixel est statique et } I_{\text{REF}}(s,t) \text{ existe} \\ 1 & \text{si le pixel est mobile} \end{cases}$$

Ce processus d'intégration temporelle tient compte des 3 éléments suivants :
- La construction de I_{REF} au pixel s n'est possible que si s est détecté comme statique (c'est à dire appartient à l'arrière plan fixe de la scène).
- La mise à jour de I_{REF} est nécessaire pour prendre en compte les variations d'illumination ou les changements de contenu du fond. Un délai de quelques (15) images est observé avant la mise à jour de $I_{\text{REF}}(s, t+1)$ dans le cas de pixels "incohérents" : ceux qui sont statiques, mais pour lesquels la différence entre $I_{\text{REF}}(s,t)$ et $I(s,t)$ est trop élevée. Ainsi, nous évitons l'intégration dans l'image de référence de bruit temporel fort et d'objets qui deviennent brièvement immobiles.

– I_{REF} est maintenue identique pour les pixels détectés comme mobiles.

4 Résultats

Les figures 3 et 4 montrent quelques résultats de segmentation. Sur les images de référence, les pixels noirs représentent des pixels pour lesquels l'image de référence n'est pas encore construite. Sur la figure 3(a), de nombreux pixels de l'image de référence ne sont pas encore disponibles, ce qui n'empêche pas une détection des objets mobiles qui s'appuie sur l'information de changements temporels. Sur la figure 3(b), on observe que des personnages presque immobiles sont maintenant détectés grâce à l'image de référence qui a eu le temps d'être complétée. La figure 4 montre que le même algorithme fonctionne également pour d'autres types de scènes, y compris des scènes intérieures.
La cadence de traitement est d'environ 6 images par seconde pour des images 288×352 avec un programme écrit en langage C non optimisé sur un PC d'entrée de gamme (700 MHz). Cette cadence est donc une estimation basse de la vitesse potentielle de l'algorithme. Plus précisément, le fait de remplacer des itérations ICM par des opérations morphologiques apporte un gain de vitesse de 45% sur cette partie de l'algorithme, qui est de loin la plus longue (70% du total).

5 Conclusion

Un algorithme d'extraction d'objets vidéo a été développé dans le contexte du projet européen `art.live`. Il utilise un modèle markovien qui prend en compte des différences inter-images et une image de référence. Seulement 2 seuils de binarisation doivent être ajustés lorsque les conditions d'acquisition changent. La cadence de traitement de 6 images par seconde est très prometteuse car elle pourrait facilement être accélérée par la réécriture de parties clé du programme en assembleur utilisant éventuellement des instructions SIMD. Bien qu'il a été développé pour le projet `art.live`, l'algorithme proposé est suffisamment générique pour être utilisé dans d'autres applications telles que la télésurveillance.

Références

[1] Aach (T.), Kaup (A.) et Mester (R.). – Statistical model-based detection in moving videos. *Signal Processing*, vol. 31, n2, Mars 1993, pp. 165–180.

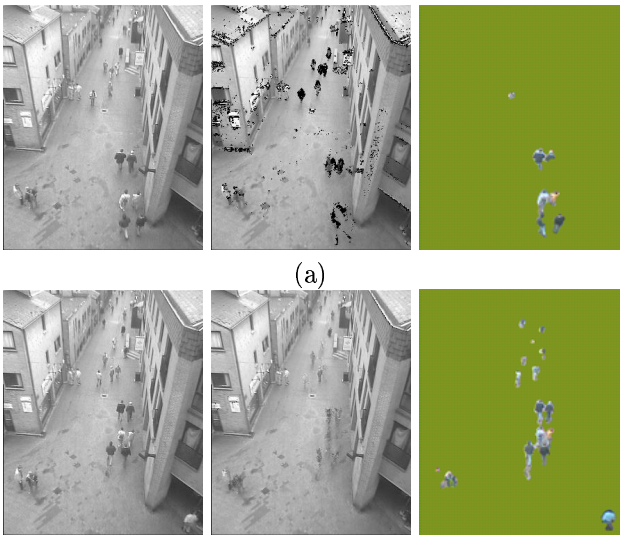
[2] Besag (J.). – On the statistical analysis of dirty pictures. *Journal Royal Statistical Society*, vol. B-48, n3, 1986, pp. 259–302.

[3] Caplier (A.), Luthon (F.) et Dumontier (C.). – Real time implementations of an mrf-based motion detection algorithm, special issue on real-time motion analysis. *Journal of Real Time Imaging*, vol. 4, n1, Fév. 1998, pp. 41–54.

[4] Cavallaro (A.) et Ebrahimi (T.). – Video objects extraction based on adaptative background and statistical change detection. *In : SPIE Electronic Imaging*. – San Jose, California, USA, Jan. 2001.

[5] Geman (S.) et Geman (D.). – Bayesian restoration of images. *IEEE Trans. Pattern Anal. and Machine Intel.*, vol. 6, n6, Nov. 1984, pp. 721–741.

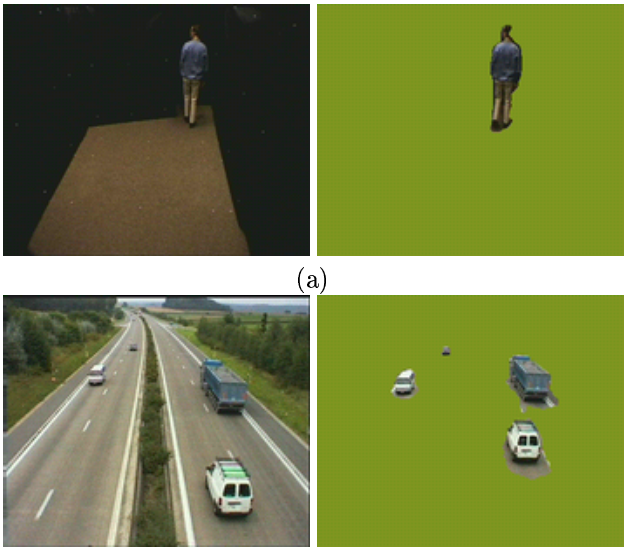
[6] Mitiche (A.) et Bouthemy (P.). – Computation and analysis of image motion : a synopsis of current problems and methods. *International Journal of Computer Vision*, vol. 19, n1, 1996, pp. 29–55.



(a)

(b)

FIG. 3 – Sur chaque ligne, de gauche à droite : image courante, image de référence et objets vidéo extraits.



(a)

(b)

FIG. 4 – (a) : scène en intérieur. (b) : trafic routier. À gauche : image, à droite : segmentation.