DETECTION DE ZONES EN MOUVEMENT DANS UNE SEQUENCE D'IMAGES SELON UNE APPROCHE MARKOVIENNE

Patrick LALANDE et Patrick BOUTHEMY

IRISA / INRIA-Rennes, Campus de Beaulieu, 35042 Rennes Cedex

RESUME

Nous abordons le problème de la détection de zones en mouvement dans une séquence d'images à travers l'analyse des variations temporelles de la distribution des intensités. En fait le problème s'étend au delà de la détection des changements temporels, à la reconstruction des projections ou masques des objets mobiles. Notre approche englobe ces deux points dans une même formalisation probabiliste introduisant une information contextuelle spatio-temporelle selon une approche Markovienne. La détection des zones mobiles est traitée comme un problème d'étiquetage statistique. Deux modèles d'étiquetage seront exposés, le premier dit par événement, le second par contenu. Des essais sur séquences réelles ont donné des résultats satisfaisants.

SUMMARY

This paper addresses the problem of motion detection in an image sequence from the variations in time of the intensity distribution. As a matter of fact, the need is not limited to change detection but encompasses the recovery of the projections of moving areas in an image. Our approach is in particular distinguished by treating conjointly these two issues, according to a probabilistic formulation. A contextual spatio-temporal information is introducted through Markovian Models. We will present two labeling models, the first one called "event-based model", the second "content-based model". Experiments with real image sequences have been carried out.

1 Introduction

Pouvoir repérer des objets en mouvement représente une étape clef en analyse de scène dynamique. Dans le cas de situations complexes (p. e. caméra en mouvement) la détermination des objets en mouvement dans la scène nécessite la reconnaissance des différents types de déplacements apparents dans l'image [4]. Cependant, pour une large classe d'applications, une simple différentiation entre zones en mouvement et zones fixes dans l'image peut suffire [1, 2, 6, 7, 8]. On est alors ramené à un problème de détection du mouvement dans l'image. Ce type de situation comprend en général une caméra fixe, pointant vers une scène 3D, où les centres d'intérêt sont les objets en mouvement. Certaines hypothèses sur le processus de formation des images, en particulier l'invariance de l'illumination sont aussi mises à profit.

Ce papier s'intéresse au problème de la détection du mouvement à partir des variations temporelles de la fonction intensité. A un objet mobile donné correspondent trois types de zones à changement temporel; la première correspond à la partie du fond découverte par l'objet, la deuxième à l'apparition de l'objet sur le fond et enfin une zone où l'objet glisse sur lui-même. Le problème est donc de ne récupérer que le masque de l'objet en mouvement. Nous proposons une méthode qui aborde conjointement le problème de la détection des changements temporels de la fonction intensité et celui de la reconstruction des zones mobiles (c.à.d. masques ou projections à un instant donné des objets en mouvement). Habituellement, ces deux étapes sont traitées l'une après l'autre, la seconde reposant souvent sur un certain nombre d'heuristiques. Une approche probabiliste est présentée ici. Des informations contextuelles en espace et temps intégrant des informations a priori sur le type de situations rencontrées sont introduites à l'aide d'une modélisation Markovienne. En fait, nous définirons deux grandes variantes de cette approche, que nous dénommerons modélisation par "événements" et modélisation par "contenus".

2 Modélisation par événement

Dans les deux cas, le problème de la reconstruction du masque des objets mobiles est abordé comme un problème d'étiquetage statistique. Décider qu'un point correspond ou non à la projection d'un objet en mouvement revient à lui attribuer telle ou telle étiquette.

Soit E le champ d'étiquettes, e une réalisation de ce champ; l'ensemble des valeurs d'étiquettes possibles est $\Omega = \{-1, 0, 1\}$. L'étiquette 0 correspond à des zones statiques, -1 et 1 à des zones en mouvement. Soit L la carte matricielle des points de l'image et p(x,y) un point de cette matrice; l'intensité au point p est notée f(p,t) ou f(x,y,t), l'image des intensités sera elle notée I_t. $e_t(p)$ représente le label du point p à l'instant t et $e_t = \{e_t(p), p \epsilon L\}$ le champ des labels. Plus généralement, si A est une partie de L, $e_t(A) = \{e_t(p), p \epsilon A\}$. Nous considérons comme observation la dérivé temporelle de la fonction intensité $\frac{\delta f}{\delta t}$. Travaillant sur des images numérisées, nous utilisons la différence finie $\tilde{f}_t(x,y) = f(x,y,t+dt) - f(x,y,t)$ comme approximation de $\frac{\delta f}{\delta t}$, dt représentant l'intervalle de temps entre deux images. Nous utilisons en pratique une version filtrée $\tilde{g}_t$ de $\tilde{f}_t$. Soit O le champ des observations et o une réalisation possible, $o_t = \{\tilde{g}_t(p), p \epsilon L\}$.

Pour résoudre le problème de labélisation, un critère de maximisation a posteriori (MAP) est pris en compte. Plus précisément, nous désirons trouver la meilleure interprétation $\hat{e}_t$, en termes de zones en mouvement, des changements temporels de la fonction intensité. Dans ce but, nous utilisons trois images successives I_{t-dt}, I_t, I_{t+dt}. De cette manière, deux champs d'observation sont obtenus o_{t-dt} et o_t, correspondant respectivement aux changements temporels entre I_t etI_{t-dt}, et entre I_{t+dt} et I_t. En effet, un seul champ d'observations n'est pas suffisant pour discerner les différentes zones créées par un objet en mouvement. En fait, nous retenons une information symbolique de changement entre t et t-dt notée $\bar{o}_{t-1}$ avec valeur dans -1,0,1. Le critère est alors le suivant (en prenant dt=1) :

$$max_{e_t} P(e_t / o_t, \bar{o}_{t-1}) \qquad (1)$$

Nous désirons prendre en compte des informations contextuelles à la fois d'un point de vue spatial et temporel. L'utilisation de modèles stochastiques, et en particulier les champs Markoviens, permet de traduire, à l'aide d'outils mathématiques appropriés, les interactions entre différents pixels voisins. De façon équivalente, nous avons recours à des distributions de Gibbs qui sont de la forme $P(x) = \frac{1}{z} exp(\frac{-U(x)}{T})$, où U, appelé fonction d'énergie, représente la somme des potentiels locaux associés aux différentes cliques. Une clique est un ensemble de points ou sites mutuellement voisins.

Le premier type d'interaction est de nature spatiale. On considère que: $P[e_t(p)/e_t(L_{/p})] = P[e_t(p)/e_t(\eta_p)]$,
où $L_{/p}$ désigne tous les points de L sauf p. Un voisinage du second ordre, une fenêtre 3x3 centrée sur p, est retenu pour η_p. Les interactions spatiales doivent exprimer les propriétés voulues du champ e_t, par exemple l'homogénéité des zones en mouvement. Ceci se traduit dans le choix des fonctions de potentiels associées aux cliques. Seules les cliques à deux éléments sont considérées : une verticale $\binom{*}{*}$, une horizontale $(**)$, deux diagonales $(*_*)$,$(_**)$. Une clique spatiale sera notée c_s. Les potentiels sont définis par :

$$\begin{cases} V_{c_s} = \beta_s & si \quad e_t(p_1) \neq e_t(p_2) \\ V_{c_s} = -\beta_s & si \quad e_t(p_1) = e_t(p_2) \end{cases} \qquad (2)$$

Ce type de potentiels favorise la continuité spatiale du champ d'étiquettes. Une carte des contours considérée comme une information déterministe peut être introduite ce qui conduit à modifier en conséquence la valeur de ces potentiels [2]. Tous ces potentiels locaux contribuent à la fonction d'énergie spatiale $W_s(e_t) = \sum_{c_s \epsilon C_s} V_{c_s}(e_t)$

Dans le modèle d'interaction temporelle sont pris en compte le champ de labels désiré e_t et le champ symbolique des changements $\bar{o}_{t-1}$. La clique temporelle est formée de deux points (p,t) et (p,t-1), c.à.d deux points ayant la même localisation spatiale mais à deux instants différents t et t-1. Cette clique sera notée c_τ. Avant de définir les potentiels associés à cette clique, certaines considérations sur la détection de zones en mouvement vont être introduites, rendant plus explicite le choix des potentiels. Si l'on suppose que les projections de l'objet sur le plan image ne se chevauchent pas entre les instants t-1, t et t+1,(c.à.d. que l'on n'est pas en présence de recouvrement), un point peut être labélisé "en mouvement" à t si la série suivante est observée : fond à t-1, objet à t, fond à t+1. En terme d'observations, nous désirons valider deux changements temporels successifs et de signes opposés, c.à.d la série d'événements (d'où le terme modélisation par événement) apparition-disparition . A l'inverse les successions objet-fond-fond ou fond-fond-objet, bien que partiellement génératrices de changements temporels, doivent être éliminées. Ceci nous conduit au choix de potentiels temporels suivant:

$$\begin{cases} V_{c_\tau} = \beta_\tau & si \quad (e_t(p) = \bar{o}_{t-1}(p) \quad OR \quad \bar{o}_{t-1}(p) = 0) \\ & \qquad\qquad\quad AND \quad e_t(p) \neq 0 \\ V_{c_\tau} = -\beta_\tau & si \quad e_t(p) = -\bar{o}_{t-1}(p) \quad OR \quad e_t(p) = 0 \end{cases} \qquad (3)$$

Le dernier cas correspondant à $e_t(p) = 0$ signifie qu'une non détection est préférée à une fausse alarme. La fonction d'énergie W_τ associée à ces potentiels temporels est donc : $W_\tau = \sum_{c_\tau \epsilon C_\tau} V_{c_\tau}(e_t, \bar{o}_{t-1})$

La dernière étape consiste à relier observation et interprétation. La relation entre le label et l'observation est modélisée par:

$$o_t(p) = \psi[e_t(p)] + n(p) \qquad (4)$$

où $\psi[e_t(p)] = m_p e_t(p)$ et n(.) est un bruit blanc gaussien centré de moyenne 0 et de variance σ^2; pour deux points p et q quelconques $\forall t$, n(p) et n(q) sont supposés indépendants. La variance σ^2 est estimée une fois au début de la séquence. Différentes stratégies sont possibles dans le choix de m_p. La première consiste à prédéfinir les valeurs de m_p. Ce modèle est facilement paramétrable mais dans certains cas irréaliste. Cependant il peut être suffisant dans une large gamme d'applications . Une seconde stratégie est actuellement en cours d'implantation. Elle consiste à estimer m_p en ligne. La fonction d'énergie W_ϵ, exprimant l'écart entre la fonction observée et l'interprétation est donnée par :

$$W_\epsilon(o_t, e_t) = \frac{1}{2\sigma^2} \sum_{p \epsilon L} [\tilde{g}_t(p) - m_p e_t(p)]^2 \qquad (5)$$

Nous pouvons donc écrire l'énergie totale du système:

$$W(e_t, o_t, \bar{o}_{t-1}) = W_s + W_\epsilon + W_\tau \qquad (6)$$

Au champ d'étiquettes le plus probable correspond la fonction d'énergie W minimum. En effet, on peut poser que $P(e_t/o_t, \bar{o}_{t-1})$ est proportionnel à $exp{-W}$. La procédure déterministe d'optimisation du critère que nous utilisons est

dérivée de celle présentée dans [5]. Elle comprend une politique de visites des sites permettant de se focaliser constamment sur les points les moins bien étiquetés, ce qui permet de réduire efficacement le nombre global d'itérations. Par ailleurs, la phase d'initialisation est réalisée via une méthode de vraisemblance. Ceci est détaillé dans [3].

Cependant, cette méthode est pleinement efficace si les projections de l'objet aux instants t et t+dt ne se recouvrent pas. Ceci revient à choisir un échantillonnage temporel dt adéquat ou à avoir une connaissance a priori sur le rapport entre la taille et la vitesse des objets recherchés.

Une modification de cette méthode permet de résoudre partiellement ce problème. Si le recouvrement est significatif, on aura tendance à seulement récupérer les contours des objets ayant un mouvement apparent. La modification apportée porte d'une part sur les étiquettes et d'autre part sur le choix des potentiels liés aux cliques. L'ensemble des étiquettes devient $\Omega = \{0, 1\}$ avec 0 pour les zones statiques, 1 sinon. Les potentiels spatiaux utilisés précédemment favorisaient les régions homogènes. Ils sont modifiés afin de prendre en compte les caractéristiques d'un contour, soit donc de favoriser la création de zones filiformes. Le potentiel temporel est défini de façon à favoriser les situations où les deux points de la clique ont la même étiquette. Toutefois ces deux méthodes, bien que donnant des résultats satisfaisants exploitables dans de nombreuses situations d'intérêt, ne permettent pas de reconstruire véritablement le masque d'un objet mobile si un fort recouvrement temporel existe. En faisant l'hypothèse que la distribution des intensités sur un objet n'est pas totalement uniforme, nous présentons dans la suite une autre façon de prendre en compte le problème de la reconstruction du masque d'un objet mobile.

3 Modélisation par contenu

Une façon différente d'aborder le problème de la reconstruction des masques des objets mobiles est de considérer l'image uniquement en fonction de son contenu: fond ou objets en mouvement. Un point de l'image appartient soit au fond: état b, soit à un objet: état a. Deux images successives I_{t-dt} et I_t sont utilisées. Elles fournissent deux types d'informations; o_t champ des observations obtenu comme précédemment par un filtrage de la dérivé temporelle de la fonction intensité mais cette fois entre t et t-dt; $\bar{o}_t$ est un champ symbolique prenant ses valeurs dans $\Omega = \{0, 1\}$ avec 1 si changement temporel, 0 sinon. Il est obtenu par un test de vraisemblance où la fonction intensité est modélisée comme étant localement une fonction linéaire bruitée [7]. Nous supposons que ce test est suffisamment sensible pour détecter les faibles variations d'intensité créées par un objet en glissement sur lui-même. Par contre, il génère de nombreuses détections parasites mais qui peuvent être éliminées par la considération d'informations contextuelles. Le problème posé est de trouver le meilleur couple (e_{t-dt}, e_t) possible, étant données les observations obtenues, soit le critère:

$$Max\ P(e_{t-dt}, e_t \,/\, o_t, \bar{o}_t) \tag{7}$$

L'optimisation du critère s'effectue par "couple glissant" c.à.d que l'on considère (e_{t-2}, e_{t-1}) puis (e_{t-1}, e_t), puis (e_t, e_{t+1}), etc...
En effet, pour complètement estimer e_t une information sur le futur est nécessaire. Ceci doit être mis en relation avec la version précédente où il était nécéssaire de considérer le triplet I_{t-dt}, I_t, I_{t+dt} dans l'estimation de $\hat{e}_t$. On a donc un retard d'une unité pour vraiment estimer $\hat{e}_t$, la premiére estimation obtenue à l'occasion du couple (e_{t-1}, e_t) servant d'initialisation pour le couple (e_t, e_{t+1}). Le champ e_t est considéré comme Markovien en espace et temps. Le voisinage spatio-temporel utilisé est constitué de deux fenêtres 3x3, centrées en un même point p(x,y) mais aux instants t-dt et t. Les potentiels spatiaux sont calculés sur chacune de ces deux fenêtres. Les cliques spatiales utilisées et leurs potentiels associés sont les mêmes que ceux définis au paragraphe 2. Nous supposons maintenant qu'un objet en mouvement entraine des variations de la fonction intensité détectable sur les trois groupes de régions évoqués en introduction. Ceci revient à supposer que les objets recherchés ne sont pas parfaitement uniformes. A un couple (a,b) ou (b,a) correspond un changement temporel important soit $\bar{o}_t(p) = 1$; on a également $\bar{o}_t(p) = 1$ pour le couple (a,a) où les changements temporels bien que faibles sont détectés. A l'opposé, le couple (b,b) n'entraine pas de changement temporel, l'information symbolique $\bar{o}_t(p)$ est donc égale à 0. Ceci permet alors de définir la fonction d'énergie temporelle:

$$W_\tau = \sum_{c_\tau \epsilon C_\tau} V_\tau(e_{t-dt}(p), e_t(p), \bar{o}_t(p)) \tag{8}$$

où les valeurs des potentiels sont données par:

$(e_{t-dt}, e_t, \bar{o}_t)$	$V_\tau(e_{t-dt}, e_t, \bar{o}_t)$
(b,b,0)	$-\beta_\tau$
(b,b,1)	$+\beta_\tau$
(a,b,0)	$+\beta_\tau$
(a,b,1)	$-\beta_\tau$
(b,a,0)	$+\beta_\tau$
(b,a,1)	$-\beta_\tau$
(a,a,0)	$+\beta_\tau$
(a,a,1)	$-\beta_\tau$

La fonction d'énergie liée à l'écart entre l'observation et l' interprétation est du même type que précédemment:

$$W_\epsilon(o_t, e_{t-dt}, e_t) = \frac{1}{2\sigma^2} \sum_{p \epsilon L} [\tilde{g}(p) - \psi(e_{t-dt}(p), e_t(p))]^2 \tag{9}$$

avec:

$$\psi(e_{t-dt}(p), e_t(p)) = \begin{cases} 0 & si \quad (e_{t-dt}(p), e_t(p)) = (b,b) \\ m_1 & si \quad (e_{t-dt}(p), e_t(p)) = (a,a) \\ m_2 & si \quad (e_{t-dt}(p), e_t(p)) = (a,b) \\ & \quad ou\ (b,a) \end{cases} \tag{10}$$

avec $0 < m_1 << m_2$. La procédure d'optimisation du critère est analogue à celle utilisée précédemment.

4 Résultats et conclusion

Des résultats sur une séquence d'images sous-marines peuvent être trouvés dans [2]. Les resultats présentés ici ont été obtenus sur une séquence de scène extérieure. La figure 1 montre une image de la séquence d'origine. La figure 2 contient l'image $\bar{o}_t$, carte des changements temporels obtenue par un test de vraisemblance. La figure 3 contient l'image

résultat où le masque de l'objet mobile (en l'occurence un piéton) est symbolisé par la tache blanche. Ces résultats ont été obtenus à l'aide de la première méthode présentée en prenant un intervalle de temps dt plus large que celui correspondant à la fréquence vidéo. Le traitement de la séquence à la fréquence originelle à l'aide de la seconde méthode est en cours.

L'approche décrite dans ce papier possède les caratéristiques suivantes. Elles permet d'aborder comme un tout la détection de changements temporels et la reconstruction des masques des objets mobiles. Elle repose sur une formalisation bien fondée des informations contextuelles spatio-temporelles. Cette modélisation du problème permet de s'adapter à différentes situations tout en conservant le même cadre d'optimisation. Enfin, elle conduit à un critère statistique autorisant une interprétation appropriée même en présence d'observations bruitées.

Des deux variantes proposées, la seconde est certainement mieux posée et plus générale. Cependant, la première garde un intérêt évident dans certains contextes d'application. En effet, celle ci reste mieux adaptée dans le cas où le recouvrement est faible (échantillonnage temporel large) et les zones recherchées uniformes. Le détecteur de changement temporel employé peut être alors à la fois simple et robuste. Cette version travaille en instantané en considérant de façon autonome trois images successives. A l'opposé, la modélisation par contenu introduit naturellement une idée de suivi temporel. Le type de détecteur de changement temporel, plus fin que le précédent, reste opérationnel quelque soit la fréquence temporelle et est plus adapté à la recherche d'objets texturés.

Bibliographie

[1] BLOSTEIN, S.D. and HUANG, T.S. A tree search algorithm for target detection in images sequences. In *Proc. IEEE Conf. CVPR, Ann Arbor*, pages 690–695, Juin 1988.

[2] BOUTHEMY, P. and LALANDE, P. Determination of apparent mobile areas in an image sequence for underwater robot navigation. In *Proc. IAPR Workshop on Computer Vision: Spec. Hardw. and Ind. Applc., Tokyo*, pages 409–412, Oct 1988.

[3] BOUTHEMY, P. and LALANDE, P. Motion detection in an image sequence using gibbs distribution. In *ICASSP, Glasgow*, may 1989.

[4] BOUTHEMY, P. and SANTILLANA RIVERO, J. A hierarchical likelihood approach for region segmentation according to motion-based criteria. In *Proc. 1st Int. Conf. on Computer Vision, London*, pages 463–467, juin 1987.

[5] CHOU, P.B. and RAMAN, R. On relaxation algorithms based on Markov random fields. *TR 212, Computer Science Dpt, Univ. of Rochester*, Juil. 1987.

[6] DONOHOE, G.W., HUSH, D.R., and AHMED,D. Change detection for target detection and classification in video sequences. In *Proc. ICASSP*, pages 1084–1087, 1988.

[7] HSU, Y. Z., NAGEL, H.-H, and REKERS, G. New likelihood test methods for change detection in image sequences. *Computer Vision, Graphics and Image Processing*, CVGIP-26:pages 73–106, 1984.

[8] WIKLUND, J. and GRANLUND, G. H. Image sequence analysis for object tracking. In *Proc. 5th Scandinavian Conf. on Image Analysis, Stokholm*, pages 641–648, Juin 1987.

Figure 1: I_t une image de la séquence (caméra fixe, piéton mobile) [ces images nous ont été fournies par le Laboratoire d'Electronique de Clermont-Ferrand et ont été numérisées au CCETT de Rennes]

Figure 2: $\bar{o}_t$, carte des changements temporels obtenue par une méthode de vraisemblance (modèle local constant pour la fonction intensité).

Figure 3: Image résultat (méthode avec modélisation par événement). Les zones mobiles validées sont surimposées en blanc.