TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

DETECTION ET SUIVI D'OBJETS PAR ANALYSE DE SEQUENCES D'IMAGES

Patrick PINEAU et Philippe ROBERT

THOMSON-CSF Laboratoires Electroniques de Rennes (LER)
Avenue de Belle Fontaine 35510 Cesson-Sévigné

RÉSUMÉ

Résumé : L'objectif du système décrit dans cet article est de détecter et de suivre des régions en mouvement dans une séquence d'images, à travers l'analyse des variations spatio-temporelles de la distribution des intensités. Pour cela, on construit les masques complets des objets mobiles, ainsi que ceux de leur ombre portée grâce à des techniques de comparaisons d'images et de contours, suivies d'un processus de relaxation avec modélisation markovienne.
La prise en compte du mouvement permet de séparer les objets appartenant à un même masque et de mettre en œuvre un processus de prédiction temporelle qui accroît la robustesse de la détection.

ABSTRACT

Abstract: The objective of the method discribed in this paper is to detect and track moving objects in image sequences through spatio-temporal changes of the luminance. For this purpose, both complete masks of the moving objects and of their shadows are built by using comparison of the luminance and of edges, followed by a relaxation process with markov modelling.
The technique includes the exploitation of motion in order to separate the objects contained in the same mask that move differently. Moreover, a temporal prediction process is introduced to increase the robustness of the segmentation.

1 Introduction

La détection de changements temporels constitue une étape importante en analyse de scène dynamique. Cependant, la variété et la diversité des techniques rencontrées dans la littérature, que ce soit dans le domaine de la détection ou de l'estimation de mouvement, s'oppose à l'existence d'une solution universelle [6,8].
L'approche envisagée ici traite des séquences d'images de scènes proches acquises avec une caméra fixe, par exemple des scènes de type trafic routier urbain. Le système développé comporte trois grandes étapes [7] :

- La détection d'objet : en entrée de cette étape on dispose de l'image courante et d'une image de référence représentant le fond de la scène observée. Une comparaison de ces deux images sur la base de la luminance et des contours conduit à une classification de chaque pixel. Cette classification repose sur un étiquetage double, chaque étiquette portant une double information (une sur les régions et une sur les contours). Cette première étape conduit à l'élaboration des masques des objets et à ceux de leur ombre portée.
- La séparation des objets connexes : chaque masque des objets peut contenir plusieurs objets en mouvement, qu'il faut alors pouvoir distinguer. Cette séparation se fera sur la base de l'estimation du mouvement des objets.
- La prédiction temporelle : pour augmenter la robustesse de la détection, chaque masque des objets est projeté dans la direction de son mouvement et constitue ainsi une estimée des masques pour l'image suivante. Cette prédiction est prise en compte par la première étape.

De même, le champ de mouvement obtenu à l'étape précédente est projeté temporellement et est utilisé pour augmenter la robustesse de la technique de séparation des objets pour l'image suivante. Cette prédiction de mouvement est prise en compte par la seconde étape.
Notons aussi qu'à la fin de cette étape, avant de traiter l'image suivante, on procède à une réactualisation de l'image de référence afin de prendre en compte les variations d'éclairage pouvant intervenir.

Dans les deux premières étapes, des techniques de relaxation avec modélisation par champs de Markov sont mises en œuvre, ce qui permet d'exploiter facilement les prédictions de l'étape 3.

2 Détection d'objet

La caméra étant fixe, une image de référence, représentant le fond de la scène observée absente de tout objet en mouvement, peut être extraite. Une simple comparaison de l'intensité lumineuse entre cette image et l'image courante permet une première classification des pixels en trois types de régions que l'on notera par la suite : fond, objet, ombre.

Une seconde classification des pixels est effectuée sur la base des contours. Après extraction des contours dans l'image courante et référence par l'opérateur de Deriche [4], ceux-ci sont comparés grâce à une technique probabiliste permettant ainsi de classer chaque point en cinq classes de contours que l'on notera : contour fond (point appartenant au fond

fixe), contour objet (point appartenant aux régions en mouvement), contour caché (contour du fond fixe caché par un objet; ces points appartiennent aux régions en mouvement), contour ombré (point appartenant au fond fixe situé dans les zones d'ombre) et non contour.
Chaque point de l'image courante $I(p,t)$ possède maintenant deux étiquettes différentes. Chaque étiquette traduit l'appartenance du point à l'une des trois zones : fond, objet et ombre. Pour rendre cette information plus robuste, nous allons fusionner ces étiquettes pour aboutir à une étiquette unique. Cela sera pleinement justifié dans le chapitre suivant lors de la relaxation.
Les deux informations à fusionner n'étant pas toujours compatibles, seules les sept étiquettes suivantes ont alors un sens : L_1=(fond,contour fond), L_2=(fond,non contour), L_3=(objet,contour objet), L_4=(objet ,contour caché), L_5=(objet,non contour), L_6=(ombre,contour ombré), L_7=(ombre,non contour).
Très peu d'incompatibilité se produit entre les deux étiquetages et lorsqu'il y en a, on constate que dans la majorité des cas, c'est l'étiquette région qui est fausse car elle est plus sensible au bruit d'image. Lorsqu'une incompatibilité intervient, seule la partie d'étiquette correspondant à la région est modifiée de façon à obtenir une des sept étiquettes ci-dessus.

Cet étiquetage constitue l'initialisation d'une procédure de relaxation avec modélisation par champs de Markov. Une telle procédure a été introduite par [5] et exploitée par la suite dans de nombreuses applications de traitement d'images [1].
Soit E le champ des étiquettes, e une réalisation de ce champ et e_p, l'étiquette au point p qui prend ses valeurs dans $L = \{L_1, ..., L_7\}$. Considérons le voisinage η_p d'ordre 2 et les cliques à deux éléments sur ce voisinage. L'optimisation de ce champ d'étiquettes est basée sur un critére de type MAP : maximisation a-posteriori de la distribution du champ des étiquettes E étant données les observations O: $\max_E P(E/O)$. Dans notre cas, deux observations sont à considérer :
O_1 : la luminance de l'image courante $I(p,t)$
O_2 : la différence entre les contours de l'image courante et ceux de l'image de référence
Le critère à maximiser devient :
$\max_E P(E/O_1, O_2)$, soit, compte tenue des règles de Bayes $\max_E P(O_1/E)P(O_2/E)P(E)$.
$P(E)$ est défini grâce aux distributions de Gibbs par :
$P(E = e) = \frac{1}{z} exp\Big(-U_c(e)\Big)$ où U_c est une fonction d'énergie définie comme la somme des potentiels sur les cliques : $U_c = \sum_{c \in C} V_c(e)$. Notons que la configuration des étiquettes la plus probable est celle ayant l'énergie U la plus faible. Les potentiels traduisent les interactions locales sur les cliques et sont définis de manière à exprimer certaines propriétés du champ des étiquettes (homogénéité par exemple).
Soient p_1 et p_2 deux points dans une clique, le potentiel sur cette clique est défini comme montré dans le tableau 1.
Ce type de potentiel favorise la continuité des étiquettes au niveau des régions et des contours et contribue donc à la construction de zones homogènes. De plus, il impose une contrainte supplémentaire qui gère les discontinuités grâce à l'étiquetage double qui prend toute son importance ici.

p_1 / p_2	L_1	L_2	L_3	L_4	L_5	L_6	L_7
L_1	-3β	$-\beta$	$+\beta$	$+\beta$	$+3\beta$	$+\beta$	$+\beta$
L_2	$-\beta$	-3β	$+\beta$	$+\beta$	$+3\beta$	$+\beta$	$-\beta$
L_3	$+\beta$	$+\beta$	-3β	$-\beta$	$-\beta$	$+\beta$	$+\beta$
L_4	$+\beta$	$+\beta$	$-\beta$	-3β	$-\beta$	$+\beta$	$+\beta$
L_5	$+3\beta$	$+3\beta$	$-\beta$	$-\beta$	-3β	$+3\beta$	$+3\beta$
L_6	$+\beta$	$+\beta$	$+\beta$	$+\beta$	$+3\beta$	-3β	$-\beta$
L_7	$+\beta$	$-\beta$	$+\beta$	$+\beta$	$+3\beta$	$-\beta$	-3β

Tableau 1

En effet, puisqu'il comporte deux informations différentes sur chaque point, il impose des contraintes fortes sur la cohabition de deux étiquettes dans une clique.
Pour définir la probabilité conditionnelle, $P(O/E)$, il est nécessaire d'établir une relation entre les observations et le champ des étiquettes :

$$O(p) = \Phi(e_p) + bruit$$

$\Phi(e_p)$ dépend de la nature de l'observation et le bruit est généralement supposé blanc, gaussien, centré de variance σ^2.
Dans ce cas, la probabilité conditionnelle devient :

$$P\Big(O(p)/e_p\Big) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{\Big(O(p) - \Phi(e_p)\Big)^2}{2\sigma^2}\right)$$

Maximiser le critère MAP revient à minimiser la fonction d'énergie a-posteriori suivante :

$$U^P = \sum_{c \in C} V_c(e) + \sum_p \left[\frac{(O_1(p) - \Phi_1(e_p))^2}{2\sigma^2}\right] + \sum_p \left[\frac{(O_2(p) - \Phi_2(e_p))^2}{2\sigma^2}\right]$$

soit $U^P = U_c + U_{o1} + U_{o2}$, avec U_c l'énergie sur les cliques, U_{o1} et U_{o1} l'énergie issue de la probabilité conditionnelle des observations O_1 et O_2 respectivement.
La fonction $\Phi_1(e_p)$ est définie par :

$$\Phi_1(e_p) = \begin{cases} I_{réf}(p) & \text{si } e_p = L_1 \text{ ou } L_2 \\ moy\_obj & \text{si } e_p = L_3 \text{ ou } L_4 \text{ ou } L_5 \\ moy\_omb & \text{si } e_p = L_6 \text{ ou } L_7 \end{cases}$$

En effet, lorsque l'étiquette indique la présence du fond fixe, la luminance de l'image courante doit être égale à la luminance de l'image de référence. Les paramètres *moy_obj* et *moy_omb* sont estimés sur des zones de taille $n \times n$ centrées autour du point p. Ils représentent la valeur moyenne de la luminance des points ayant l'étiquette objet et ombre respectivement dans la zone $n \times n$ considérée. Ce calcul est effectué en utilisant les étiquettes de l'étiquetage initial. Lorsqu'aucune étiquette objet et/ou ombre n'est présente dans la zone $n \times n$, *moy_obj* et/ou *moy_omb* sont fixés à des valeurs arbitraires.

Soit $e_{préf} \in \{contour, non\ contour\}$ l'étiquette contour dans la référence. La fonction $\Phi_2(e_p, e_{préf})$ est alors définie par :

$$\Phi_2(e_p, e_{préf}) = \begin{cases} CM_1 & \text{si } (e_p, e_{préf}) = (L_1, \text{contour}) \\ CM_2 & \text{si } (e_p, e_{préf}) = (L_3, \text{contour}) \\ CM_3 & \text{si } (e_p, e_{préf}) = (L_4, \text{contour}) \\ CM_4 & \text{si } (e_p, e_{préf}) = (L_6, \text{contour}) \\ CM_5 & \text{si } (e_p, e_{préf}) = (L_2, \text{non contour}) \\ CM_6 & \text{si } (e_p, e_{préf}) = (L_3, \text{non contour}) \\ CM_7 & \text{si } (e_p, e_{préf}) = (L_5, \text{non contour}) \\ CM_8 & \text{si } (e_p, e_{préf}) = (L_7, \text{non contour}) \end{cases}$$

Les CM_i sont estimés par apprentissage sur quelques images. Remarquons que toutes les combinaisons possibles $(e_p, e_{préf})$ ne sont pas représentées dans Φ_2 afin d'éviter les incompatibilités entre la classe de contour attribuée aux points de la source et la classe de contour des points de la référence.
La procédure déterministe d'optimisation du critère que nous utilisons est dérivée de celle présentée dans [2]. Elle comprend une stratégie de visite des sites permettant de se focaliser sur les points les moins bien étiquetés, ce qui permet de réduire le nombre d'itération.

3 Séparation de objets connexes

Des résultats précédents, nous pouvons extraire les régions fond, ombre et objet. Cependant, rien ne nous permet de distinguer la présence de différents objets à l'intérieur d'une même région objet. Seule la prise en compte du mouvement va nous permettre de séparer les objets connexes.
Comme la plupart des techniques d'estimation de mouvement fournissent des résultats trop bruités pour permettre de trouver les frontières délimitant les différents objets à l'intérieur d'une même région, nous avons adopté une méthode comportant quatre phases distinctes.

- Premièrement on procède à une estimation de mouvement en tous points d'image avec l'estimateur décrit dans [3], puis à la sélection des vecteurs majoritaires dans ce champ.
- Ces vecteurs majoritaires sont assignés à chaque point des régions objet sur la base de la $|DFD|$ (*Displaced Frame Difference*) minimum.
- Cette attribution de mouvement est ensuite optimisée grâce à une technique de relaxation avec modélisation markovienne du champ de mouvement. Pour chaque région objet, les paramètres de cette modélisation sont les suivants :

L'observation est la luminance de l'image courante $I(p,t)$. Les étiquettes sont les vecteurs de mouvement $D^i, i = 1, n$ sélectionnés dans l'histogramme des mouvements et attribués en chaque point. L'étiquette en un point p est notée D_p et prend ses valeurs dans $L = (D^1, \ldots, D^n)$.
Nous utilisons un voisinage d'ordre 2 et les cliques à deux éléments sur ce voisinage.
L'énergie issue de la probabilité conditionnelle de l'observation est :

$$U_o = \sum_p \left[\frac{\Big(I(p,t) - I(p - D_p, t - dt)\Big)^2}{4\sigma^2} \right]$$

La fonction traduisant la compatibilité de deux vecteurs de mouvement dans la clique favorise donc la présence du même mouvement sur les points de la clique si ceux-ci ont la même étiquette e_p. Si les points ont une étiquette e_p différente, on n'est pas en mesure d'influencer l'attribution du mouvement (en particulier, on ne sait pas si un contour voisin d'un non contour appartient ou non au même objet). Le potentiel sur les cliques temporelles est alors défini par :

$$V_c(D_p) = \begin{cases} -\beta & \text{si } D_{p1} = D_{p2} \text{ et } e_{p1} = e_{p2} \\ +\beta & \text{si } D_{p1} \neq D_{p2} \text{ et } e_{p1} = e_{p2} \\ 0 & \text{si } e_{p1} \neq e_{p2} \end{cases}$$

L'énergie spatiale sur les cliques est : $U_c = \sum_{c \in C} V_c(D)$ La technique de relaxation déterministe est la même que celle utilisée dans le paragraphe précédent.

4 Prédiction temporelle

Le lien temporel existant naturellement entre les images successives de la séquence est exploité afin d'augmenter la robustesse de la détection et d'obtenir ainsi une meilleure détection et une meilleure stabilité des masques au cours du temps.
La prédiction temporelle intervient à deux niveaux différents :

- Sur la détection d'objet : la détection de l'étape 1 va être enrichie par une image prédite temporellement issue de la projection de chaque masque des objets mobiles dans le sens de leur mouvement. Cette information est introduite dans la relaxation pour l'optimisation du champ des sept étiquettes de l'étape 1 en tant que clique temporelle. L'énergie totale à minimiser devient alors : $U^P = U_c + U_{o1} + U_{o2} + U_{c_\tau}$ où U_{c_τ} est l'énergie sur les cliques temporelle c_τ : $U_{c_\tau} = \sum_{c_\tau \in C_\tau} V_{c_\tau}(e)$ Cette énergie traduit la compatibilité entre les étiquettes de l'image courante $e_p \in \{L_1, ..., L7\}$ et les étiquettes de l'image de prédiction $e_\tau \in \{objet, fond\}$. Elle doit être définie de manière à favoriser la continuité temporelle de l'étiquetage. Le potentiel est alors défini par :

$$V_{c_\tau} = \begin{cases} -\beta' & \text{si } e_\tau = \textit{objet mobile} \\ & \text{et } e_p = \{L_3, \text{ ou } L_4, \text{ ou } L_5\} \\ +\beta' & \text{si } e_\tau = \textit{objet mobile} \\ & \text{et } e_p = \{L_1, \text{ ou } L_2, \text{ ou } L_6, \text{ ou } L_7\} \\ 0 & \text{si } e_\tau = fond \end{cases}$$

- Sur la séparation des objets connexes : l'attribution du mouvement de l'étape 2 va être enrichie par un champ de mouvement prédit, issu de la projection temporelle du champ de mouvement précédent. Ce champ prédit intervient aussi au niveau de la relaxation pour l'optimisation du champ de mouvement en tant que clique temporelle. L'énergie sur les cliques temporelles traduit la compatibilité entre le mouvement issu de l'attribution initiale

$D \in \{D^1_{t+dt}, \ldots, D^n_{t+dt}\}$ et le mouvement prédit $D_\tau \in \{D^1_t, \ldots, D^m_t\}$. Le mouvement variant peu d'une image à l'autre, cette énergie doit alors être définie de manière à favoriser cette faible variation.
Nous avons alors défini une distance entre la valeur prédite et chaque valeur candidate. L'estimée de mouvement candidate favorisée sera celle qui fournit la plus petite distance. Soient DX_τ et DY_τ le mouvement horizontal et vertical prédit en un point p, et DX^i et DY^i, $i = 1, \ldots, n$ les mouvements horizontaux et verticaux candidats sur le masque objet, alors la distance entre D_τ et D^i est définie par :

$$DIST_i = \sqrt{(DX_\tau - DX^i)^2 + (DY_\tau - DY^i)^2}$$

et le potentiel sur les cliques temporelles par :

$$V_{c_\tau} = \begin{cases} -\beta' & \text{si } DIST_i < DIST_j \quad i \neq j, \quad i,j = 1, \ldots, n \\ 0 & \text{sinon} \end{cases}$$

5 Résultats et conclusions

Cet algorithme a été testé sur des images de trafic routier et les résultats sont montrés figure 1. Le fond est représenté en noir, les objets en blanc et leur ombre en gris. La détection est de bonne qualité et les masques des objets sont restitués fidèlement.
La méthode développée permet de s'adapter à différents types de scène à fond fixe et reste relativement robuste pour les raisons suivantes :
• Les régions d'ombre sont prises en compte de manière explicite évitant ainsi les fausses détections.
• La prise en compte des contours permet d'approximer au mieux la silhouette des objets.
• La prédiction temporelle agit à deux niveaux : d'une part, individuellement, chaque prédiction améliore les résultats et d'autre part, elles interagissent du fait des boucles imbriquées.

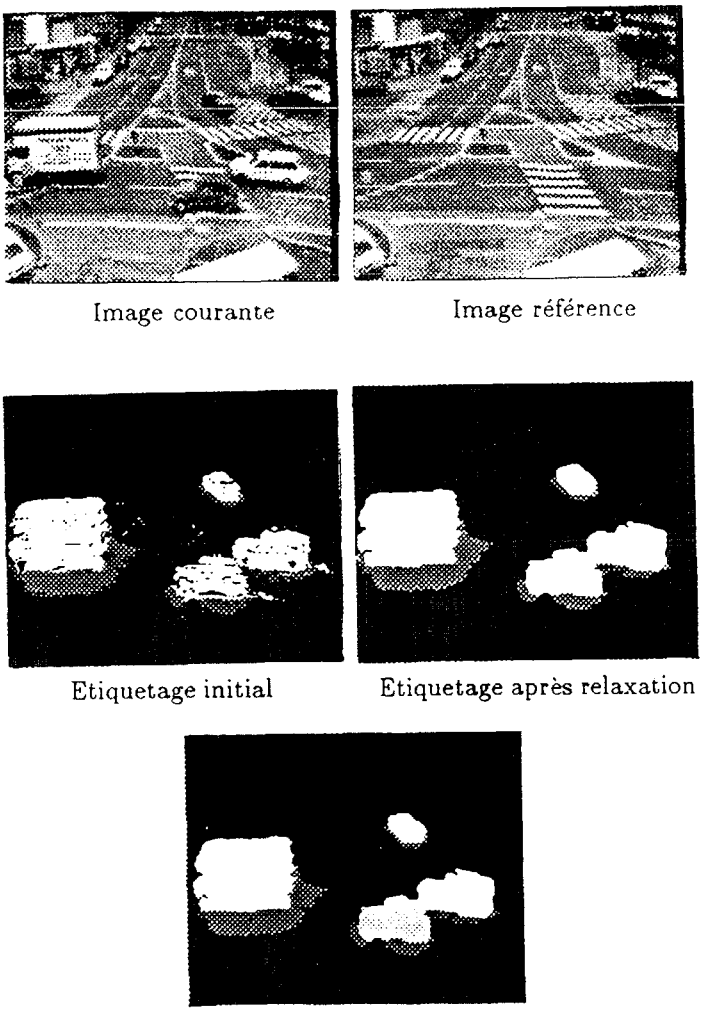

Figure 1 : Résultats

Bibliographie

[1] CHELLAPPA R. "Two-dimensional discrete gaussian Markov random field models for image processing and analysis" *SPIE, Vol 1075, Digital Image Processing Applications, 1989.*
[2] CHOU P.B., RAMAN R. "On relaxation algorithms based on markov random fields" *TR 212, Computer Science Department, University of Rochester, New-York, July 1987.*
[3] CHUPEAU B., ROBERT P., PECOT M., GUILLOTEL P. "Multiscale motion estimation" *Proc Worshop on Advanced Matching in Vision and Artificial Inteligence, Munich, Juin 90.*
[4] DERICHE R. "Using Canny's criteria to derive a recursively implemented optimal edge detector" *Int. Journal of Computer Vision, 1987.*
[5] GEMAN S., GEMAN D. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images" *IEEE Trans. on PAMI, Vol. 6, N°6, 1984.*
[6] NAGEL H.H. "Image sequences. Ten (octal) years. From phenomenology towards a theoretical foundation" *Proc 8th Int. Conf. on Pattern Recognition, Paris, Oct. 1986.*
[7] PINEAU P. "Détection et suivi d'objets par analyse de séquences d'images" *Thèse de l'Université de Rennes I, 1991.(à paraître)*
[8] WIKLUND J., GRANLUND G.H. "Image sequence analysis for object tracking" *Proc. 5th Scandinavian Conf. on Image Analysis, Stokholm, June 1987.*