Détection des zones de mouvement et régularisation: Application à la vidéo surveillance

Lionel CARMINATI, Jenny BENOIS-PINEAU

LaBRI UMR CNRS 5800
351 Cours de la libération
33405 Talence (France)
{lionel.carminati, benois-p}@labri.fr

Résumé – L'objectif de notre étude consiste à concevoir un système de vidéo surveillance orienté objet qui affiche les activités d'un site en détectant et suivant un objet d'intérêt. Cet objet connu est supposé être animé d'un mouvement propre différent de celui de la caméra -supposée statique dans notre cas-. Le problème consiste donc à localiser la zone d'intérêt, c'est à dire la zone de mouvement et de la rendre la plus nette et lisse possible au sens spatio-temporel pour pouvoir servir à la recherche d'un ou plusieurs objets d'intérêt. La méthode proposée actuellement s'exécute en temps réel (25 images par seconde) sur des vidéos de résolution CIF.

Abstract – The goal of this work is to create an object oriented video surveillance system that monitors activity at a site over extended periods of time by detecting motion activities in a site. The paper deals with detection of moving objects by modelling pixel grey level distribution along the time. The detection of moving objects is based on learning and update of background pixel distributions. The choice of appropriate mixture's component for a given pixel is performed by likelihood maximization. An original Markov regularization is proposed to smooth detection. The method performs in real time on CIF resolution video and low cost commercial hardware.

Le problème de détection des zones de mouvement avec une caméra fixe dans la vidéo a été largement étudié dans la littérature [1, 2, 3] . Récemment la modélisation à base de mélange de lois Gaussiennes est devenue très populaire grâce à certains travaux fondamentaux [1, 4, 5, 6]. Appliqué au problème de la télésurveillannce, ces modèles montrent de bonnes performances pour des conditions d'éclairage variables au cours du temps.

Le modèle à base de mlange de Gaussienne est devenue très populaire du fait que l'intensité d'un pixel peut être modélisé de façon efficace par un mélange de lois Gaussiennes à condition de supposer un bruit d'acquisition non corrélé et de faibles changements de luminosité. Les auteurs de [1, 5] proposent ainsi de modéliser le fond de la scène -par opposition aux objets en mouvement- par un mélange de lois Gaussiennes des valeurs RGB d'un pixel. Grimson propose également un schéma de mise à jour dans lequel les pixels sont confrontés au modèle existant afin de permettre leur classification : "fond"/"objet en mouvement".

Considérant le cadre du modèle par mélange de lois Gaussiennes, nous proposons dans ce papier une méthode originale de détection de mouvement appliquée à des contextes de vidéo surveillance. Le processus est découpé en trois étapes : initialisation du mélange, classification de chaque nouveau pixel et mise à jour du modèle en fonction de la classification. La méthode proposée diffère des autres travaux par l'apprentissage et le réentrainement que l'on suggère. Une régularisation markovienne est également étudiée afin d'améliorer le taux de détection tout en ajoutant une cohérence à la fois spatiale et temporelle. Enfin contrairement aux travaux précédents [1, 4] oú les auteurs proposent un simple seuillage statistique sur le poids des distributions, une règle de décision basée sur le maximum de vraisemblance nous permet de déterminer la meilleure distribution Gaussienne parmi le mélange. Dans la méthode proposée le nombre initial de Gaussiennes par pixel, obtenu lors de l'entrainement, et donc supposé optimal, reste inchangé.

Ce papier est organisé de la façon suivante : A travers la Section 1 nous présentons le modèle par mélange de lois Gaussiennes et la règle de décision que nous proposons pour classifier tous les nouveaux pixels en "fond" / "objet en mouvement". Notre schéma de mis à jour ainsi que les équations permettant de mettre à jour le modèle en temps réel sont présentés également dans cette même section. La Section 2 présente quand à elle le schéma de régularisation de type Markovienne du champ d'étiquettes. Les résultas sur des vidéos extraites de notre corpus de télésurveillance sont décrits à la section 3. La conclusion et les perspectives seront finalement présentées à la Section 4

1 Modèle à mélange de lois Gaussiennes et sa mise à jour

Considérant les valeurs de niveaux de gris de chaque pixel de l'image comme un processus stochastique indépendant, nous supposons que la distribution suivie par les valeurs de luminance d'un pixel peut être modélisée par un mélange de lois Gaussiennes si la probabilité de la luminance x_t à l'instant t est définie comme suit

$$P(x_t) = \sum_{i=1}^{K} w_{i,t} * \eta(x_t|\, \mu_{i,t}, \sigma^2_{i,t}) \qquad (1)$$

oú $w_{i,t}$ est le poids , $\mu_{i,t}$ et $\sigma^2_{i,t}$ sont respectivement la moyenne

et l'écart type de la $i$ème Gaussienne à l'instant t. Le but de notre travail consiste donc à estimer la densité de probabilité $P(x_t)$ pour chaque pixel et pour chaque instant t. Pour ce faire nous proposons une approche en 3 étapes : la première, dite d'initialisation, estime le nombre optimal K de Gaussiennes ainsi que leurs paramètres dans le mélange durant un intervalle de temps donné grâce à l'algorithme ISODATA [7] sur l'ensemble d'apprentissage X, des valeurs de $\{x_1, ...x_n\}$. Cet algorithme, basé sur le fameux K-Means, fusionne et découpe les clusters suivant des seuils de compacité et d'éloignement définis par l'utilisateur. Remarquons que par la suite le nombre de Gaussiennes dans le mélange n'est pas remis en cause.

La deuxième étape de détection des pixels en mouvement consiste en la vérification des hypothèses statistiques de l'appartenance de la valeur courante de l'intensité du pixel à la distribution de probabilité ou non.

Comme les conditions d'éclairage évolue au fil du temps, les paramètres du modèle d'intensité d'un pixel doit être mis à jour. Nous avons divisé ce processus en deux étapes : classification des nouvelles valeurs du pixel et mis à jour du modèle correspondant suivant la dècision précédente. La premiére étape détermine, pour chaque pixel, si son intensité lumineuse suit ou non la distribuau méltion existante c'est à dire si le pixel est "fond" ou pas. Grimson[1, 8] propose de vérifier chaque nouveau pixel avec les K distributions existantes. Si sa valeur RGB est comprise entre 2.5 fois l'écart-type de la distribution, le pixel est considéré comme appartenant à la distribution.

Dans notre étude une solution plus élégante est proposée par maximisation de la vraissemblance. Elle est réalisée de la façon suivante : Pour un exemple donné x_t nous maximisons la vraissemblance du paramètre η_i du mélange $P(x_t) = \sum_{i=1}^{K} w_{i,t}.\eta(x_t, \mu_{i,t}, \sigma^2_{i,t})$. Dans ce cas nous considérons la vraissemblance du paramètre η_i conditionnellement à l'ensemble du mélange Gaussien. Considérons une partition complète de l'espace des hypothèses, X, définie par $B = \{H_1, ..., H_i, ..., H_K\}$ et assoçions chaque Gaussienne η_i avec H_i. Grâce au thèorème de Bayes, nous avons $P(H_i/B) = P(H_i.B)/P(B)$. Supposant que H_i, $i = 1, .., K$ forme une partition complète de B nous avons $P(H_i.B) = P(H_i)$.

Considérons la densité de probabilité $p(x_t \in H_i/B)$ telle que

$$p(x_t \in H_i/B) = \frac{w_i \eta_i}{\sum_{k=1}^{K} w_k \eta_k} \tag{2}$$

La vraissemblance conditionnelle de la $i^{\text{ème}}$ Gaussienne du mélange pour l'exemple x_t sera alors

$$l(x_t) = \frac{w_i \eta_i}{\sum_{k=1}^{K} w_k \eta_k} \tag{3}$$

Suivant le processus usuel de prise de décision nous maximisons la log-vraissemblance $L_i = log(l(x))$. Nous avons alors

$$L_i = log\,(\frac{w_i.\eta_i}{\sum_{k=1}^{K} w_k \eta_k}) = log\,(w_i \eta_i) - log\,(\sum_{k=1}^{K} w_k \eta_k) \tag{4}$$

comme $log\,(\sum_{k=1}^{K} w_k \eta_k)$ est le même pour tous les L_i nous devons alors maximiser $log\,(w_i \eta_i)$ ce qui est finalement équivalent à chercher η_i^* avec sa correspondance $\Theta_i^* = (w_i^*, \mu_i^*, \sigma_i^{2*})$ tel que

$$\Theta_i^* = \underset{\Theta_i=(w_i,\mu_i,\sigma_i^2)}{argmax} \left(log\, w_i - \left(\frac{log\, \sigma_i^2}{2} + \frac{(x-\mu_i)^2}{2\sigma_i^2} \right) \right) \tag{5}$$

En considérant la vraisemblance de chaque Gaussienne dans le mélange associé au pixel par rapport à la mesure courante x, nous déterminons la "meilleure" Gaussienne au sein du mélange grâce à la règle de décision ci dessus. Ainsi si le maximum - partie droite de l'équation- est au dessus d'un certain seuil, la nouvelle valeur x_t du pixel sera considérée comme appartenant au fond. Dans le cas contraire, on supposera que le pixel appartient à l'objet en mouvement.

Une fois la classification des pixels effectuée, la dernière étape consiste à mettre à jour le modèle. Nous proposons un réentrainement adapté en fonction du résultat de classification précédent. Si l'intensité du pixel suit une distribution déjà existante -c'est à dire si le pixel fait partie du fond- alors les poids et les paramètres (μ_i, σ_i^2) du mélange Gaussien seront mis à jour de la même façon que dans [1], c'est à dire :

Les poids seront ajustés en considérant l'équation suivante

$$w_{i,t} = (1-\alpha) w_{i,t-1} + \alpha v(w_k|x_t) \tag{6}$$

oú $\alpha \in [0, 1]$ est le taux d'apprentissage fixé par Grimson, $v(w_k|x_t)$ est égale à 1 si $k = i^*$, i.e. $\eta_k(\mu_k, \sigma_k^2)$ est la meilleure correspondance pour x_t, et 0 sinon. Moyenne et écart-type sont mis à jour en considérant

$$\mu_{i,t} = (1-\alpha)\mu_{i,t-1} + \rho x_t \ , \tag{7}$$

$$\sigma^2_{i,t} = (1-\alpha)\sigma^2_{i,t-1} + \rho(x_t - \mu_{i,t})^2 \tag{8}$$

avec $\rho = \alpha \eta(x_t|\mu_{i,t}, \sigma^2_{i,t})$. En pratique, et suivant les travaux de Grimson, une valeur de α aux alentours de 0.1 donne des résultats convaincant au cours du temps.

Si la valeur d'intensité du pixel n'appartient à aucune des distributions déjà existantes pendant une période de temps fixé, nous proposons de relancer la phase d'initialisation sur l'ensemble des pixels constitué du pixel courant et des pixels qui ont été détecté comme étant un objet en mouvement pendant cette même période. Contrairement aux autres travaux [1] , cette méthode permet une bien meilleure réactivité face aux changements rapides de luminosité ou lorsqu'un objet en mouvement s'arrête de bouger.

A chaque itération de temps t, les pixels qui ont été comparés aux Gaussiennes $\eta_i(t)$ sont étiquetés b si ils appartiennent au fond a sinon. Après l'étape de détection de mouvement, on obtient une carte de champ d'étiquette que nous allons régulariser par approche markovienne.

2 Régularisation Markovienne temps réel

Du fait du bruit intrinsèque de la caméra et du mouvement des petits objets -feuilles, ...- il peut rester parfois des fausses détections indésirables. De plus on remarque que certains des objets résultants ne sont pas tout à fait complets. Une approche type régularisation Markovienne permet donc de résoudre ce problème [9, 2] . Généralement, on modélise la différence entre deux images par seulement une Gaussienne, ce qui permet de formuler une fonction d'énergie englobant les termes d'attaches

aux données et de régularisation des champs d'étiquettes. Dans le cas d'un mélange de lois Gaussiennes, une telle formulation est impossible du fait que les distributions des intensités de pixels suivent une loi additive [2]. En considérant seulement les champs d'étiquettes (a, b) résultant de l'étape de détection, on peut quand même proposer la régularisation suivante non pas sur les valeurs de pixels mais plutt sur la carte d'étiquette obtenue après classification.

Soit O l'ensemble d'observations tel que $O = \{o_1, ..., o_n\}$, $E = \{e_1, ..., e_n\}$ un ensemble d'étiquettes prenant ses valeurs dans $\{a, b\}$. Nous dénoterons dès lors $\{E = e\}$ l'état particulier tel que $\{E_{s1} = e_{s1}, ..., E_{sn} = e_{sn}\}$. De plus, considérons $v(n)$ un voisinage du site s_n et V un système de voisinage. Le fameux théorème d'Hammersley et Clifford établit que $P(E = e) = \frac{1}{Z} exp(-\sum V(e))$. Le problème consiste alors à maximiser la probabilité conditionnelle $P(O = o|E = e).P(E = e)$. En supposant le problème au sens du Maximum A Posteriori -M.A.P-, nous définissons la fonction d'énergie $U(e)$ comme la décomposition de deux termes :

$$U(e) = \sum_{c \in C} V(e) = V(e_s, e_t) + V(e, o) \tag{9}$$

$V(e_s, e_t)$ est le terme de régularisation avec $< s, t >$ une clique binaire de c sur le voisinage V. $V(e, o)$ est le terme d'attache aux données. Nous exprimons ces deux termes de la façon suivante

$$V(e_s, e_t) = \begin{cases} 0 & if \quad e_s = e_t \\ \beta & sinon \end{cases} = \beta(1 - \delta_{e_s, e_t}) \tag{10}$$

avec δ le symbole de Kronecker. $V(e, o)$ exprime les dépendances temporelles des étiquettes en introduisant Φ tel que

$$\Phi(e_{t-dt}(s), e_t(s)) = \begin{cases} 0 & if \quad (e_{t-dt}(s), e_t(s)) = (b, b) \\ m_1 & if \quad (e_{t-dt}(s), e_t(s)) = (a, a) \\ m_2 & sinon \end{cases} \tag{11}$$

La maximisation de la fonction d'énergie s'effectue ensuite grâce à l'algorithme "Iterated Conditional Modes" (I.C.M.) [10] .

3 Résultats

Nous avons testé notre méthode sur deux types de corpus. Le premier contient 10 séquences de 5000 images de locaux à surveiller. Le deuxiéme est composé de 5 séquences de vidéo surveillance prise à l'extérieur contenant 2500 images chacune. Chaque image est de résolution CIF. Quelques résultats sont présentés sur la Figure (1). Du fait des variations brutales de luminosité et de la position de la source lumineuse, chaque type de corpus présente des contraintes différentes : le premier implique des changements rapides de luminosité et un objet peut traverser rapidement la scène. La détection doit par conséquent être rapide et la mise à jour également.

Dans le deuxième corpus les changements sont moins rapides et le processus d'apprentissage doit tre plus lent pour converger vers la distribution de "fond" idéale tout en restant le plus efficace possible au fil du temps.

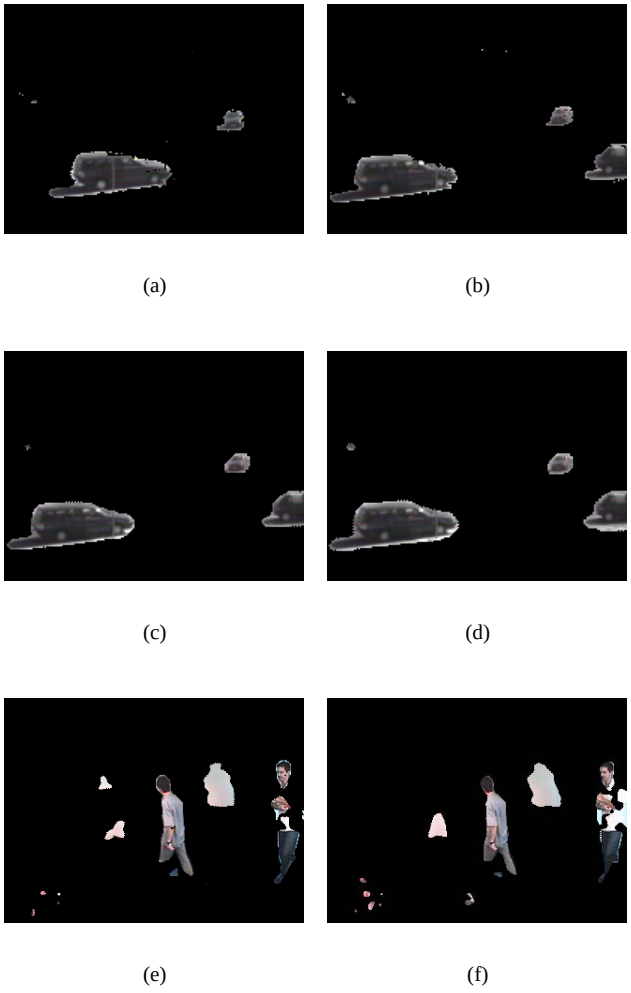

FIG. 1 – Exemple de masque de mouvement extraits du corpus de vidéo surveillance. La ligne du haut présente la détection de mouvement sans régularisation. La seconde ligne présente les masques de mouvement après régularisation Markovienne. L'algorithme détecte les feuilles en haut à droite, mais la régularisation efface ces fausses détections. La dernière ligne présente les résultats pour une scène d'intérieur avec une régularisation type markov.

4 Conclusion

Ces travaux s'inscrivent dans la continuté de ceux de Grimson. L'approche que nous proposons consiste à détecter les objets en mouvement pour un système de télésurveillance. Nous avons développé une méthode originale basée sur la modélisation par mélange de lois Gaussiennes sur une fenêtre temporelle des valeurs de luminosité. En travaillant uniquement sur les valeurs de luminosité et grâce à une classification par maximisation de la vraissemblance, le processus que nous proposons permet d'être déployer sur des systèmes temps-rèel avec des architectures banalisées. Une régularisation type Markov ajoute au masque de mouvement une cohérence spatio-temporelle, permettant ainsi d'éliminer les fausses détections.

Références

[1] W. Grimson, C. Stauffer, R. Romano, L.Lee, Using adaptive tracking to classify and monitor activities in a site, IEEE CVPR 1998 (1998) 22–29.

[2] P. Lalande, P. Bouthemy, A statistical approach to the detection and tracking of moving objects in an image sequence, 5th European Signal Processing Conference EUSIPCO 90.

[3] L. Carminati, J. Pineau, M. Gelgon, Human detection and tracking fro video surveillance applications in low density environment, SPIE VCIP'2003 SPIE 0277 -786X 5150 (2003) 51–60.

[4] C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking, the Artificial Intelligence Laboratory, Massachussets Institute of Technology Cambridge, MA02139 (1998).

[5] P. Kaewtrakulpong, R.Bowden, An improved adaptative background mixture model for real-time tracking with shadow detection, 2nd European Workshop on Advanced Video Based Surveillance Systems.

[6] K. Karmann, A. Brandt, Detection and tracking of moving objects by adaptative background extraction, Proceedings of the 6th Scandinavian Conference on Image Analysis (1989) 1051–1058.

[7] G. Ball, D. Hall, A clustering technique for summarizing multivariate data, in : Behavioral Science, Volume 12, 1967, pp. 153–155.

[8] C. Stauffer, W. Eric, W. Grimson, Learning patterns of activity using real-time tracking, IEEE PAMI, Volume 22 (8) (2000) 747–757.

[9] S. Geman, D.Geman, Stochastic relaxation, gibbs distributions and the bayesian restoration of images, IEEE PAMI 1984 Vol.6.

[10] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society (1974) 36 :192–236 Series B.