Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Étiquetage statistique d'un graphe de régions pour la détection d'objets mobiles dans des séquences d'images couleur

Ronan Fablet[1], Patrick Bouthemy[2], Marc Gelgon[2]
[1]IRISA/CNRS [2]IRISA/INRIA
Campus universitaire de Beaulieu 35042 Rennes Cedex, France
email : Ronan.Fablet@irisa.fr

Résumé – Nous cherchons à détecter des objets mobiles dans des séquences d'images couleur acquises avec une caméra mobile. Ce problème est essentiel pour de nombreuses applications. Afin de retrouver plus efficacement les frontières de mouvement, nous exploitons une partition spatiale très fine de l'image issue d'une segmentation markovienne sur un critère de couleur. Nous introduisons une modélisation sur un graphe de régions dans un contexte markovien. Nous formulons ce problème de détection des objets mobiles comme l'étiquetage de ce graphe en régions conformes et non-conformes au mouvement dominant. Nous avons validé notre approche sur des exemples de séquences réelles.

Abstract – We aim at detecting moving objects in color image sequences acquired with a mobile camera. This issue is of key importance in many application fields. To recover accurately and efficiently motion boundaries, we make use of a fine spatial image partition delivered by a first MRF-based color segmentation stage. We introduce a region-level graph modeling embedded in a Markovian framework both to achieve this color-based spatial segmentation and to detect moving objects in the scene viewed by a mobile camera. This is stated as the labeling of this graph into regions conforming or not to the dominant image motion. The method is validated on real image sequences.

1 Introduction

L'extraction des objets mobiles dans des séquences d'images constitue une tâche fondamentale d'un grand nombre d'applications : navigation d'un robot, suivi de cibles, surveillance vidéo, indexation vidéo,... Généralement, ces dernières ne requièrent pas une segmentation complète de l'image au sens du mouvement mais uniquement l'extraction des entités pertinentes. La détection des éléments en mouvement dans la scène observée reste une tâche difficile si la caméra est elle-même mobile.

La plupart des méthodes proposées pour la détection du mouvement reposent sur des techniques de classification au niveau des pixels exploitant une mesure locale liée au mouvement apparent dans l'images telle que la DFD (Displaced Frame Difference) ou la vitesse normale. La phase de classification en zones statiques et mobiles des pixels de l'image s'appuie sur des procédures de seuillage, [6, 7, 11], ou des approches bayésiennes, [9, 12, 14]. De plus, certaines approches procèdent itérativement au niveau des pixels puis au niveau des régions, [14]. L'extraction d'une partition spatiale initiale de l'image a aussi été largement exploitée pour la segmentation au sens du mouvement dans des séquences d'images que ce soit à partir de critères de mouvement, [15], ou d'informations d'intensité, de texture et de couleur, [1, 4, 13]. Ces dernières offrent une meilleure précision quant à la localisation des frontières de mouvement, qui correspondent généralement aussi à des frontières au sens de l'intensité, de la texture ou de la couleur. A partir de cette segmentation initiale, un modèle paramétrique 2D de mouvement est associé à chaque région spatiale, et la phase de segmentation au sens du mouvement consiste à réaliser une fusion de régions. Elle peut exploiter des techniques de classification dans l'espace des paramètres de mouvement, [13], ou des approches bayésiennes telles que l'utilisation du critère MDL, [15], ou des techniques markoviennes d'étiquetage contextuel sur un graphe de régions, [4]. L'une des limitations de ces approches réside dans le fait qu'elles ne peuvent exploiter des partitions spatiales très fines afin de garantir la pertinence de l'estimation de modèles de mouvement paramétriques. Elles risquent donc de perdre certaines frontières de mouvement.

Dans cet article, nous décrivons une approche se plaçant au niveau région visant à directement détecter les objets mobiles dans la scène à partir d'une séquence d'images couleur acquises par une caméra mobile. Pour cela, de manière similaire à [4], nous exploitons une partition préliminaire de l'image au sens de la couleur. Cependant, nous n'associons pas de modèles de mouvement paramétriques à chaque région spatiale. Nous estimons uniquement une représentation du mouvement global dominant dans l'image, et intégrons des informations de mouvement locales pour juger de la pertinence dans chaque région spatiale du mouvement dominant estimé.

Nous procédons en trois étapes. Tout d'abord, nous estimons un modèle affine 2D global représentant le mouvement dominant. Puis, un graphe, dont les noeuds correspondent aux régions spatiales, est déduit d'une segmentation de l'image au sens de la couleur. Dans un troisième temps, nous introduisons un modèle markovien pour réa-

liser un étiquetage des noeuds du graphe en deux classes exprimant le fait qu'une région est conforme ou non au mouvement dominant. Si ce dernier correspond au mouvement induit par la caméra, ce qui est généralement le cas, les régions déclarées non-conformes contiennent les objets mobiles dans la scène. Nous avons porté une attention particulière à la modélisation de ce problème par une fonction d'énergie exploitant d'une part des mesures locales de mouvement appropriées et incorporant d'autre part une information contextuelle au niveau des régions. De plus, cette approche permet d'utiliser une segmentation spatiale très fine conduisant à une localisation précise des frontières des objets mobiles dans l'image.

Cet article est organisé de la manière suivante. En section 2, la modélisation markovienne envisagée pour réaliser un étiquetage statistique sur un graphe de régions sera détaillée. Nous décrirons ensuite les différents modules de notre méthode de détection de mouvement dans la section 3. Enfin, la section 4 présente des résultats sur des séquences réelles et nous concluons en section 5.

2 Étiquetage statistique de graphe de régions

Nous supposons disposer d'une première partition spatiale de l'image à traiter. Nous cherchons à fusionner certaines régions sur des critères de couleur ou de mouvement. Dans cette optique, nous considérons une approche par étiquetage statistique du graphe d'adjacence $\mathcal{G}$ défini par l'ensemble $\mathcal{N}$ des régions spatiales et l'ensemble $\mathcal{A}$ des arcs reliant deux régions connectées dans la partition spatiale, [4, 14]. L'approche bayésienne envisagée repose sur la définition d'un modèle markovien sur ce graphe de régions, chaque site N étant un noeud du graphe. De plus, nous déduisons de l'ensemble des arcs $\mathcal{A}$ un système de voisinage de cliques binaires. En exploitant un critère du MAP et l'équivalence entre champs markoviens et gibbsiens, [5], la phase de fusion de régions se ramène à déterminer la configuration d'étiquetage $\hat{e}$ qui vérifie :

$$\hat{e} = \arg\min_{e} U(e, o) \tag{1}$$

où $U(e,o) = U^a(e,o) + U^b(e)$ avec o représentant l'ensemble des observations considérées aux noeuds du graphe, U^a l'énergie d'attache aux données et U^b le terme de régularisation. Ces deux fonctions d'énergie se décomposent en somme de potentiels locaux :

$$\begin{cases} U^a(e,o) = \displaystyle\sum_{N \in \mathcal{N}} V^a(e_N, o_N) \\ \\ U^b(e) = \displaystyle\sum_{(N_1,N_2) \in \mathcal{A}} V^b(e_{N_1}, e_{N_2}) \end{cases} \tag{2}$$

Le potentiel de régularisation V^b tend à favoriser des étiquettes identiques pour deux régions voisines, et prend aussi en compte le «degré» d'adjacence à travers le calcul de deux attributs géométriques :

$$V^b(e_{N_1}, e_{N_2}) = -\beta \cdot \frac{\alpha_{N_1N_2}}{\alpha_{N_1N_2} + D_{N_1N_2}} \cdot \delta(e_{N_1} - e_{N_2}) \tag{3}$$

où β est une constante, $\alpha_{N_1N_2}$ la longueur de la frontière commune aux régions N_1 et N_2, et $D_{N_1N_2}$ la distance entre les centres de gravité des deux régions.

Dans la suite, nous exploiterons cette approche pour réaliser une fusion de régions au sens de la couleur (sous-section 3.2) et la détection des régions mobiles ou plus précisément non-conformes au mouvement dominant (sous-section 3.3).

3 Détection d'objets mobiles

3.1 Estimation du mouvement

La première étape de la méthode de détection de mouvement que nous avons développée, consiste à estimer le mouvement global dominant, représenté par un modèle affine 2D, entre deux images. Le vecteur vitesse $\overrightarrow{w_\Theta}(s)$ en un point s issu du mouvement affine paramétré par Θ est donné par : $\overrightarrow{w_\Theta}(s) = (a_1 + a_2x + a_3y, a_4 + a_5x + a_6y)^T$, avec $s = (x, y)$ et $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]^T$. L'estimation du modèle est réalisée par une méthode robuste incrémentale et multirésolution, décrite dans [8]. Elle consiste à poser le problème de minimisation suivant :

$$\widehat{\Theta} = \arg\min_{\Theta} \sum_{s} \rho\left(DFD(s, \Theta)\right) \tag{4}$$

où $DFD(s,\Theta) = I_{t+1}(s + \overrightarrow{w_\Theta}(s)) - I_t(s)$, I_t est la fonction intensité dans l'image à l'instant t et $\rho()$ la fonction "biweight" de Tukey. L'estimateur robuste permet de ne pas être sensible aux mouvements secondaires dus aux objets mobiles dans la scène. La solution itérative résulte de linéarisations successives de l'expression $DFD(s,\Theta)$ associé à l'estimé courant de Θ et revient à suivre un schéma des moindres-carrés pondérés itérés à chaque étape.

3.2 Segmentation couleur

L'information de couleur permet de localiser précisément les frontières des objets dans des scènes réelles, [1, 4]. Afin d'extraire une partition initiale de l'image au sens de la couleur, nous exploitons une approche markovienne associée à une modélisation gaussienne de la distribution de couleur sur chaque région. De plus, afin de réduire le nombre de régions de couleur, nous considérons à la fois une régularisation au niveau des pixels et une étape de fusion de régions.

La procédure envisagée au niveau des pixels consiste à itérativement estimer le modèle gaussien associé à chaque label en utilisant les moments empiriques, puis à affiner les frontières de chaque région par une relaxation markovienne. Pour cela, nous considérons un ensemble de modèles gaussiens, et nous exploitons une modélisation markovienne du champ d'étiquettes e, correspondant aux numéros des régions de couleur homogène, sur la grille des pixels de l'image. Nous utilisons un critère du MAP, et nous définissons des potentiels locaux v^a et v^b relatifs aux termes d'attache aux données et de régularisation. En chaque site, l'observation est fournie par la composante couleur c exprimée dans l'espace décrit dans [10]. Elle est donnée par $c = (c_1, c_2, c_2)$, où $c_1 = r - v$, $c_2 = 2b - r - g$

et $c_3 = r + g + b$ avec (r, g, b) désignant les composantes (rouge,vert,bleu). Le potentiel d'attache aux données v^a en un site s est alors défini par l'expression :

$$v^a(e_s, c_s) = \eta_s(M_{e_s}, \Sigma_{e_s}) \tag{5}$$

où (M_{e_s}, Σ_{e_s}) est le modèle gaussien associé au label e_s avec $\eta_s(M_{e_s}, \Sigma_{e_s}) = (c(s) - M_{e_s})^t \, \Sigma_{e_s}^{-1} \, (c(s) - M_{e_s})$. Par ailleurs, le terme v^b favorise l'homogénéité spatiale :

$$v^b(e_r, e_s) = \mu(1 - \delta(e_r - e_s)) \tag{6}$$

où μ est une constante positive et (e_r, e_s) forme une clique pour le système de voisinage du second-ordre.

Dans une deuxième étape, l'approche présentée en section 2 est exploitée pour regrouper les régions qui présentent des similarités en termes de couleur. Ainsi, le potentiel d'attache aux données V^a_{coul} quantifie la capacité d'un modèle gaussien, associé à un label λ, à décrire la distribution de couleur relative à un noeud du graphe N. Au noeud N, nous comparons les modèles relatifs à l'ensemble des étiquettes des noeuds voisins. Ensuite, après affectation du meilleur label $\hat{\lambda}$, le modèle associé à $\hat{\lambda}$ est mis à jour. Par ailleurs, nous introduisons une information supplémentaire dans le terme de régularisation (équation 3), en le pondérant par un coefficient lié au contraste de couleur sur la frontière des deux régions considérées.

La procédure de segmentation au sens de la couleur est appliquée dans un cadre multi-échelle. Nous commençons par assigner une étiquette différente à chaque bloc de l'échelle la plus grossière de la pyramide. Ensuite, nous itérons à chaque échelle de plus en plus fine les procédures de régularisation au niveau pixel et de regroupement de régions. Dans les deux cas, la minimisation de la fonction d'énergie markovienne est réalisée par un algorithme du type ICM modifié, [2]. Comme aucun a priori n'est fixé sur le nombre de régions, cet algorithme est non supervisé, ce qui permet d'appréhender une large gamme de situations. De plus, afin d'extraire l'ensemble des frontières de mouvement qui sont supposées correspondre à des contours de couleur, nous considérons des segmentations spatiales très fines (typiquement, jusqu'à de 50 pixels par région).

3.3 Détection du mouvement

La détection du mouvement consiste à réaliser un étiquetage binaire de la partition obtenue au sens de la couleur en terme de régions conformes ou non au modèle de mouvement dominant estimé $\widehat{\Theta}$. Pour ce faire, nous exploitons la procédure d'étiquetage statistique décrite en section 2. Le potentiel d'attache aux données V^a_{mvt} vise à évaluer la pertinence locale du modèle de mouvement dominant estimé $\widehat{\Theta}$. L'observation o_N au niveau d'une région R_N est un ensemble $(\epsilon_s)_{s \in R_N}$ de mesures locales liées au mouvement, calculées en chaque point s de la région R_N. Cette quantité est aussi utilisée dans des approches pour la détection du mouvement se situant au niveau pixel, [7, 9]. Dans notre étude, nous mettons en outre à profit l'information de couleur. Nous considérons donc un vecteur d'observations au niveau pixel $\epsilon_s = (\epsilon^i_s)$:

$$\epsilon^i_s = \frac{\sum_{p \in \mathcal{W}(s)} |DFD^i(s, \widehat{\Theta})| \cdot \|\nabla I^i(p)\|}{\max\left(G^2_m, \sum_{p \in \mathcal{W}(s)} \|\nabla I^i(p)\|^2\right)} \tag{7}$$

où i fait référence à la composante couleur c_i et $\mathcal{W}(s)$ une fenêtre 3×3 autour de s et la constante G_m traduit le niveau de bruit dans l'image. Dans [9], des bornes d'interprétation, l^i_s et L^i_s, de la quantité ϵ^i_s, vis à vis du mouvement résiduel après compensation du mouvement dominant sont déduites de la distribution des gradients spatiaux de l'intensité pour la composante couleur c_i telles que :

$$\begin{cases} \text{si } \epsilon_s < l^i_s(\Delta) \quad \text{alors } \|\overrightarrow{w}(s) - \overrightarrow{w_{\widehat{\Theta}}}(s)\| < \Delta \\ \\ \text{si } \epsilon_s > L^i_s(\Delta) \quad \text{alors } \|\overrightarrow{w}(s) - \overrightarrow{w_{\widehat{\Theta}}}(s)\| > \Delta \end{cases} \tag{8}$$

où Δ désigne l'amplitude minimale du mouvement résiduel que l'on cherche à détecter et $\overrightarrow{w}(s)$ le mouvement réel (inconnu) au site s. En exploitant ces bornes, le potentiel V^a_{mvt} pour une région R_N correspondant à un noeud N du graphe s'exprime comme suit :

$$\begin{cases} V^a_{mvt}(conf) = \displaystyle\sum_{s \in R_N} \frac{\max_i \left[A(\epsilon^i_s, l^i_s(\Delta))\right]}{|R_N|} \\ \\ V^a_{mvt}(nconf) = \displaystyle\sum_{s \in R_N} \frac{\max_i \left[1 - A(\epsilon^i_s, L^i_s(\Delta))\right]}{|R_N|} \end{cases} \tag{9}$$

où α est une constante, $conf$ et $nconf$ correspondent respectivement aux étiquettes «conforme» et «non-conforme» et la fonction A représente une version continue d'un échelon unité.

Le nombre de noeuds dans le graphe étant petit (comparé à la taille d'une image), nous exploitons une procédure de minimisation de type HCF, [2]. La carte initiale de détection est obtenue par la minimisation de la fonction d'énergie réduite au seul terme d'attache aux données.

4 Résultats

Les expérimentations ont été menées avec les valeurs suivantes des paramètres introduits : $\mu = 0.045$, $\beta_{coul} = 0.1$ pour la segmentation au sens de la couleur ; $G_m = 15.0$, $\beta_{mvt} = 100.0$ et $\Delta = 1.0$ pour l'étiquetage du mouvement sur le graphe de régions. Le facteur Δ (pouvoir de discrimination dans la détection du mouvement) est typiquement une valeur définie par l'utilisateur et a de façon claire une incidence sur le nombre de régions détectées comme non-conformes au mouvement dominant. Étant donné sa signification physique très claire, il est facile de le fixer suivant les résultats que l'on souhaite obtenir, ce qui confère une flexibilité intéressante à notre méthode.

Dans un premier exemple, montré en Fig.1($a-b$), la caméra effectue un mouvement vers le haut et vers la droite, la voiture s'avançant vers la droite. La méthode de détection exploitant couleur et mouvement permet d'extraire des frontières de mouvement assez précises en dépit de la présence d'effets d'illumination. Le second exemple, présenté à la Fig.1($c-d$), est une scène dynamique complexe : la caméra suit un joueur de tennis. Cette scène comporte des mouvements non rigides qui sont des situations difficiles à appréhender. À nouveau, la segmentation initiale au sens de la couleur permet de localiser efficacement les

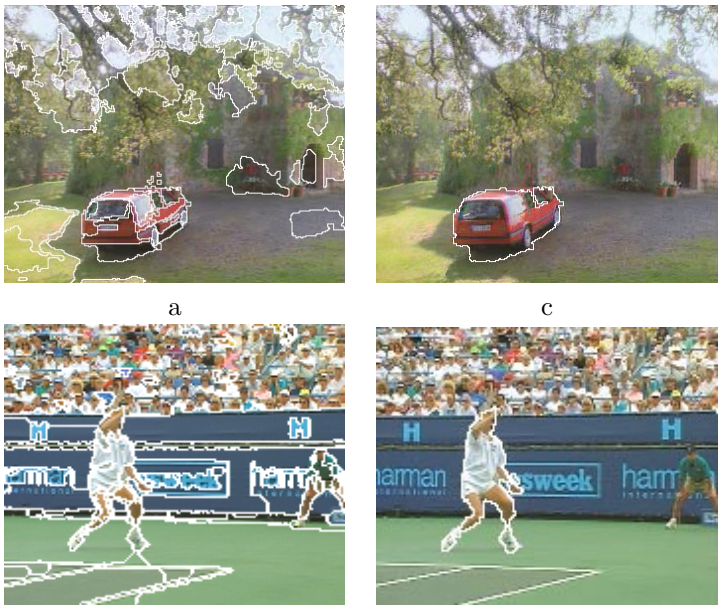

FIG. 1: *Résultats d'extraction d'objets mobiles pour les séquences "Ajax" (a-b), "Stefan" (c-d) : frontières spatiales au sens de la couleur (a-c), contours des objets mobiles détectés (b-d). La séquence "Ajax" a été fournie par l'INA (Institut National de l'Audiovisuel, Département Innovation).*

frontières de mouvement du corps et le schéma de détection du mouvement s'avère fiable et précis.

Le temps de calcul de la phase de segmentation spatiale est de l'ordre de la minute pour des images de taille 512×512, alors que l'étape de détection du mouvement proprement dite nécessite quelques secondes pour un graphe associé à une centaine de régions (sur une station de type Sun Creator 360MHZ).

5 Conclusion

Nous avons décrit une méthode de détection des objets mobiles dans des séquences d'images couleur acquises avec une caméra mobile. Nous réalisons un étiquetage contextuel du graphe d'adjacence de régions, issues d'une partition de l'image sur un critère de couleur, en terme de zones conformes ou non au mouvement dominant dans l'image, représenté par un modèle de mouvement 2D paramétrique.

Comme le montrent les résultats obtenus, la localisation des frontières de mouvement se trouve être particulièrement améliorée par la prise en compte du critère spatial de couleur. De plus, notre approche au niveau région considère une information contextuelle de plus haut niveau que pour les méthodes raisonnant au niveau des pixels, et permet donc de se rapprocher de la notion d'"objet".

Nos prochains travaux viseront tout d'abord à exploiter notre méthode de détection du mouvement pour suivre des objets mobiles dans une scène. D'autre part, cette approche pourra aussi être combinée aux travaux présentés dans [3] afin de s'affranchir des effets de perspective en présence de fortes variations de profondeur dans la scène.

Références

[1] Y. Altunbasak, P. Ehran Eren et A. Murat Tekalp. – Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, vol. 60, n° 1, jan. 1998, pp. 13–23.

[2] P.B. Chou et C.M. Brown. – The theory and practice of Bayesian image modeling. *Int. Journal of Computer Vision*, vol. 4, 1990, pp. 185–210.

[3] G. Csurka et P. Bouthemy. – Direct identification of moving objects and background from 2D motion models. *In : Proc. 7th IEEE Int. Conf. on Computer Vision, ICCV'99*. – Kerkyra, Greece, sept. 1999.

[4] M. Gelgon et P. Bouthemy. – A region-level motion-based graph representation and labeling for tracking a spatial image region. *Pattern Recognition*, 1999. – À paraître.

[5] S. Geman et D. Geman. – Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, n° 6, 1984, pp. 721–741.

[6] M. Irani et P. Anandan. – A unified approach to moving object detection in 2D and 3D scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, n° 6, juin 1998, pp. 577–589.

[7] M. Irani, B. Rousso et S. Peleg. – Detecting and tracking multiple moving objects using temporal integration. *In : Proc. 2nd Eur. Conf. on Computer Vision, ECCV'92*. – Santa Margherita, mai 1992.

[8] J.M. Odobez et P. Bouthemy. – Robust multiresolution estimation of parametric motion models. *Jal of Visual Communication and Image Representation*, vol. 6, n° 4, dec. 1995, pp. 348–365.

[9] J.M. Odobez et P. Bouthemy. – Separation of moving regions from background in an image sequence acquired with a mobile camera. *In : Video Data Compression for Multimedia Computing*, chap. 8, pp. 295–311. – H. H. Li, S. Sun, and H. Derin, eds, Kluwer Academic Publisher edition, 1997.

[10] M.J. Swain et D. Ballard. – Color indexing. *Int. Journal of Computer Vision*, vol. 7, n° 1, 1991.

[11] W.B. Thompson, P. Lechleider et E.R. Stuck. – Detecting moving objects using rigidity constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, n° 2, fev. 1993, pp. 162–165.

[12] P.H.S. Torr et D.W. Murray. – Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, vol. 11, n° 4, mai 1993.

[13] J.Y.A. Wang et E.H. Adelson. – Representing moving images with layers. *IEEE Trans. on Image Processing*, vol. 3, n° 5, sept. 1994, pp. 625–638.

[14] W. Xiong et C. Graffigne. – A hierarchical method for detection of moving objects. *In : Proc. 1st IEEE Int. Conf. on Image Processing, ICIP'94*, pp. 795–799. – Austin, nov. 1994.

[15] H. Zheng et D. Blostein. – Motion-based object segmentation and estimation using the MDL principle. *IEEE Trans. on Image Processing*, vol. 4, n° 9, sept. 1995, pp. 1223–1235.