Détection d'événements dans les vidéos à l'aide de modèles probabilistes de mouvement

Gwenaëlle Piriou[1], Nathalie Peyrard[1], Patrick Bouthemy[1], et Jian-Feng Yao[2]

[1] IRISA-INRIA Rennes, Projet Vista,
Campus universitaire de Beaulieu, 35042 Rennes cedex, France.
Tél : int+33 2 99 84 71 00, Fax: int+ 33 2 99 84 71 71
{Gwenaelle.Piriou, Nathalie.Peyrard, Patrick.Bouthemy}@irisa.fr

[2] IRMAR,
Campus universitaire de Beaulieu, 35042 Rennes cedex, France.
Tél : int+33 2 23 63 69, Fax: int+ 33 2 23 67 90
Jian-Feng.Yao@univ-rennes1.fr

Résumé Nous présentons une méthode originale de détection d'événements pertinents dans une vidéo, basée sur l'analyse du contenu dynamique et ne nécessitant pas de segmentation de mouvement préalable. Pour cela, nous exploitons des mesures locales de mouvement de bas niveaux reliées aux vitesses normales et immédiatement calculables pour tout type de vidéo, quels que soient son genre et son contenu. Une originalité de notre approche est d'appréhender séparément le mouvement de la caméra (en fait, le mouvement dominant) et le mouvement de la scène (en fait, le mouvement résiduel) dans une séquence. En effet, ces deux composantes de mouvement apportent des informations différentes, mais complémentaires qu'il est pertinent d'explicitement prendre en compte pour la classification ou la détection d'événements. Par exemple, dans une vidéo de sport, la caméra suit généralement le joueur ou l'athlète en action et par conséquent apporte principalement le mouvement global (souvent une translation globale). Le mouvement résiduel apporte quant à lui l'information sur l'activité réelle de la scène (notamment les gestes du joueur). Le mouvement résiduel est spécifié à partir des mesures locales de mouvement obtenues après compensation du mouvement dominant et directement calculées à partir des dérivées spatio-temporelles des intensités. Ces mesures sont ensuite exploitées dans un cadre statistique et un modèle probabiliste en est alors proposé. Un modèle probabiliste est également introduit pour traiter le mouvement de la caméra. Nous appliquons ce cadre statistique à la détection d'événements pertinents dans une vidéo selon un schéma à deux étapes. La première étape consiste en une présélection de segments candidats parmi les segments successifs de la vidéo traitée, selon deux groupes, représentant respectivement le "contenu dynamique important" et le "contenu dynamique négligeable". La seconde étape est une étape de classification selon le critère du maximum de vraisemblance pour reconnaître les événements pertinents (en termes de contenu dynamique) parmi les segments retenus lors de la première étape. Un tel procédé en deux temps nous permet de restreindre la phase de reconnaissance à un ensemble approprié et limité de classes afin de rendre la classification plus robuste et plus efficace. Des résultats expérimentaux sur des vidéos de sport sont décrits et montrent le bon niveau de performance de la méthode proposée.

Mots clés Mesures de mouvement, Modèles probabilistes, Détection d'événements dans les vidéos.

[1] Ce travail a été en partie financé par la Région Bretagne. Nous remercions également l'INA qui nous a fourni le corpus vidéo utilisé pour les expérimentations. Par ailleurs, pour des raisons de droits d'auteurs, nous ne sommes pas autorisés à inclure des reproductions des images de ce corpus.

1 Introduction

Un des challenges actuels dans le domaine de la vision par ordinateur est de pouvoir approcher le "contenu sémantique" de documents vidéos. Cela peut concerner des tâches comme la création de résumés de vidéos, l'indexation vidéo ou la surveillance de scènes. La principale difficulté réside dans la détection d'événements de nature plutôt sémantique à partir d'informations de bas niveau extraites des images. Les caractéristiques d'un événement doivent être exprimées en termes de primitives vidéo (couleur, texture, mouvement, forme ...) suffisamment discriminantes.
Plusieurs approches ont été développées exploitant différentes sortes de primitives vidéo. Dans [9], la sélection d'images-clefs est basée sur les composantes de couleur des pixels. Les auteurs de [8] se sont intéressés à la classification de séquences vidéos en différents genres en introduisant des modèles statistiques sur les éléments structurant une vidéo. Récemment, dans [3], une méthode de classification basée sur les SVMs ("support vector machines"), utilisant un descripteur de "texture de mouvement" a été présentée. La combinaison des informations extraites des bandes image et audio est également exploitée, par exemple dans [4], pour la création de résumé de vidéo.
Comme la variation du contenu dynamique est un puissant indicateur de l'apparition d'un événement, l'analyse du mouvement dans les séquences d'images est largement utilisée pour la segmentation des vidéos en plages significatives, ou pour la reconnaissance d'événements particuliers. Une caractérisation efficace du mouvement peut être dérivée des champs denses de vitesses, comme dans [7] pour la détection d'activité domestique. Dans [11], les auteurs utilisent des mesures spatio-temporelles très simples, à savoir des histogrammes des dérivées spatiales et temporelle de l'intensité, pour repérer des événements temporels. Dans [10], une représentation en composantes principales de paramètres de mouvement (tels que translation, rotation, ...), apprise d'un ensemble d'exemples est introduite et est appliquée à la reconnaissance de mouvements humains particuliers, une segmentation initiale du corps humain étant supposée disponible.
Nous nous intéressons dans ce papier au problème de la reconnaissance d'événements prédéfinis, à partir de l'information de mouvement. La méthode définie est générale et implique des calculs simples, puisque nous exploitons des mesures de mouvements de bas niveau, apportant cependant une information de mouvement plus élaborée que celle utilisée dans [11]. Nous traitons séparément le mouvement dominant supposé dû au déplacement de la caméra et le mouvement résiduel relatif au mouvement de la scène. Ces deux composantes du mouvement sont ensuite représentées par des modèles probabilistes appropriés décrits dans les sections 2 et 3, puis sont recombinées pour la tâche de classification. L'algorithme de détection d'événement est décrit dans la section 4. Il fonctionne en deux étapes, nécessite une phase d'apprentissage et sélectionne les segments vidéos intéressants selon le critère du maximum de vraisemblance. Des expérimentations sur des vidéos de sport ont été menées et sont décrites dans la section 5.

2 Modélisation probabiliste du mouvement de la scène

La méthode proposée pour la reconnaissance d'événements dynamiques significatifs s'appuie sur une modélisation probabiliste du mouvement contenu dans la vidéo. Pour assurer

généralité et efficacité, ce cadre exploite seulement des mesures locales de mouvement calculables directement à partir des dérivées spatio-temporelles de l'intensité. Il est possible de caractériser le mouvement perçu dans la vidéo comme proposé dans [2], en calculant des quantités locales correspondant aux moyennes pondérées d'amplitudes de vitesses normales. Cependant, le mouvement perçu est en fait la résultante de deux sources de mouvement : le mouvement de la caméra et le mouvement de la scène. Une information plus élaborée et plus utile peut en effet être récupérée en traitant ces deux types de mouvement séparément plutôt qu'en considérant seulement leur somme. Ainsi, nous estimons puis compensons le mouvement dominant entre images successives de la séquence vidéo, afin de calculer des mesures locales de mouvement correspondant seulement au mouvement résiduel. A chaque instant t, un modèle de mouvement 2D paramétrique (représenté par le vecteur de paramètres θ ; une fois estimé, il sera noté $\hat{\theta}$) représentant le mouvement dominant dans l'image est estimé comme expliqué à la section 3. Les mesures locales de mouvement $v_{res}(p,t)$ sont alors définies comme la moyenne pondérée par le carré de la norme du gradient spatial d'intensité, des amplitudes de vitesses normales résiduelles. Ces dernières sont en fait déduites de la différence d'image déplacée fournie par le mouvement dominant estimé et notée $DFD_{\hat{\theta}_t}$. On obtient :

$$v_{res}(p,t) = \frac{\sum_{q\in\mathcal{F}(p)} \|\nabla I(q)\|.|DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q\in\mathcal{F}(q)} \|\nabla I(q)\|^2\right)}, \tag{1}$$

où $DFD_{\hat{\theta}_t}(q) = I(q + \overrightarrow{w}_{\hat{\theta}_t}(q), t+1) - I(q,t)$ avec $\overrightarrow{w}_{\hat{\theta}_t}(q)$ le vecteur de vitesse fourni au pixel q par le modèle de mouvement estimé. $\mathcal{F}(p)$ est une fenêtre spatiale au point p. $\nabla I(q,t)$ est le gradient spatial de l'intensité lumineuse au pixel q et à l'instant t. η^2 est une constante prédéterminée relative au niveau de bruit.
En calculant ces quantités sur différentes séquences vidéos, les histogrammes se sont avérés être relativement proches d'une distribution gaussienne, tronquée sur les valeurs positives puisque par définition $v_{res}(p,t) \geq 0$, complétée d'un pic en zéro. Par conséquent, nous modélisons la distribution des mesures locales de mouvement par un modèle de mélange spécifique de densité :

$$f_{v_{res}}(x) = \beta\delta_0(x) + (1-\beta)\phi(x;0,\sigma^2)\mathbb{1}_{x>0} \tag{2}$$

où β est le poids du mélange, δ_0 est la fonction de Dirac en 0 ($\delta_0(x) = 1$ si $x = 0$ et $\delta_0(x) = 0$ sinon) et $\phi(x;0,\sigma^2)$ est la fonction de densité gaussienne centrée et de variance σ^2. Les paramètres β et σ^2 sont estimés selon le critère du maximum de vraisemblance (MV). Afin de capter également l'évolution de l'information de mouvement au cours du temps, nous considérons les mesures d'ordre 2 au sens d'une accélération que sont les contrastes temporels Δv_{res} des mesures locales de mouvement. Ces contrastes sont définis comme la différence temporelle des variable v_{res} données par (1) :

$$\Delta v_{res}(p,t) = v_{res}(p,t+1) - v_{res}(p,t). \tag{3}$$

Ils sont encore modélisés par un modèle de mélange d'une mesure de Dirac en 0 et d'une distribution gaussienne centrée, mais cette fois non tronquée. Le poids du mélange et la variance de la distribution gaussienne sont, comme précédemment, évalués par MV. Le modèle probabiliste global du mouvement résiduel est ensuite défini comme le produit des deux modèles introduits :

$$P_{\mathcal{M}_{res}}(v_{res}, \Delta v_{res}) = P(v_{res}).P(\Delta v_{res}) \tag{4}$$

3 Modélisation probabiliste du mouvement de la caméra

Nous devons élaborer un modèle probabiliste du mouvement de la caméra dans le but de le combiner avec le modèle probabiliste du mouvement résiduel lors du processus de reconnaissance. Premièrement, le mouvement dominant dans l'image est représenté par un modèle affine 2D déterministe de la forme suivante :

$$\overrightarrow{w}_\theta(p) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix}, \tag{5}$$

où $\theta = (a_i, i = 1, \ldots, 6)$ est le vecteur de paramètres du modèle et $p = (x, y)$ est un point de l'image. Ce modèle de mouvement relativement simple peut appréhender différents mouvements de caméra tels que les panoramiques, les zooms, les travellings. Les paramètres θ du modèle de mouvement sont estimés par l'algorithme robuste multi-résolution décrit dans [5]. A ce stade, il pourrait être possible de caractériser directement le mouvement de la caméra par le vecteur θ de paramètres et de représenter sa distribution sur la séquence par un modèle probabiliste. La principale difficulté dans ce cas, est de proposer un modèle probabiliste pertinent. En effet, si la distribution des deux paramètres de translation a_1 et a_4 peut être facilement décrite (ces paramètres sont supposés être constants entre deux changements de mouvement de la caméra de telle façon qu'un mélange gaussien pourrait raisonnablement être utilisé), la tâche se complique pour ce qui est des autres paramètres qui ne sont pas constants et qui ne sont pas non plus de la même nature. Pour cette raison, nous proposons de construire la carte des vecteurs de vitesses associés au mouvement dominant en chaque pixel de l'image, une fois le modèle de mouvement affine estimé, et d'exploiter ces mesures comme un histogramme 2D. Plus précisément, à chaque instant t, les paramètres de mouvement θ_t du modèle du mouvement de la caméra (5) sont estimés et les vecteurs $\overrightarrow{w}_{\hat{\theta}_t}(p)$ sont calculés en chaque point p du support de l'image. Les composantes horizontales et verticales des $\overrightarrow{w}_{\hat{\theta}_t}(p)$ sont ensuite finement quantifiées afin de construire l'histogramme 2D empirique de leur distribution sur la séquence. Finalement, cet histogramme est représenté par un modèle de mélange de distributions gaussiennes. Le nombre de composantes dans le mélange est déterminé par le critère ICL (Integrated Completed Likelihood, [1]) et les paramètres du modèle sont ensuite estimés par l'algorithme EM.

4 Algorithme de détection d'événement

Nous exploitons maintenant les modèles probabilistes de mouvement introduits pour une tâche de détection d'événement. Nous supposons que les vidéos à traiter sont préalablement segmentées en plages temporelles homogènes, par exemple selon la méthode décrite dans [6]. L'algorithme de détection d'événement est composé de deux étapes. La première permet de trier les segments vidéos en deux groupes. Le premier groupe rassemble les segments susceptibles de contenir des événements pertinents, le second est formé des segments à ignorer. Typiquement, si nous considérons des vidéos de sport, nous tentons de distinguer les segments de phases de jeux "Play" des autres segments "No play". Cette étape est uniquement basée sur le mouvement résiduel qui explique le mouvement de la scène. Pour cela, un modèle de mouvement est appris pour chacun des deux groupes lors d'une phase d'apprentissage.

	P	NP
P	16	3
NP	1	10

	SP	RP	LC	RC
SP	2	0	0	0
RP	0	2	0	0
LC	0	1	4	1
RC	0	0	0	6
NP	0	1	0	0

Fig. 1. A gauche : matrice de confusion associée au tri (étape 1 de l'algorithme). A droite : matrice de confusion correspondant à l'étape 2 de l'algorithme. Les lignes forment la vérité terrain. Les colonnes contiennent les étiquettes attribuées.

Ensuite, le tri consiste à attribuer l'étiquette "Play" ou "No play" à chacun des segments de la vidéo traitée, selon le critère du maximum de vraisemblance. En pratique, du fait de la grande diversité de contenu dans les segments "Play" et "No play" de certaines vidéos, il peut être nécessaire d'apprendre plusieurs modèles par groupe.
La seconde étape du schéma proposé consiste à reconnaître quelques événements spécifiques parmi les segments préalablement sélectionnés. Contrairement à la première étape, les deux types d'information (mouvement résiduel et mouvement de caméra) sont requis puisque leur combinaison permet de caractériser plus précisément un événement particulier. Pour un genre donné de contenu vidéo, une phase d'apprentissage est réalisée. Lors de cette phase, un modèle de mouvement résiduel $\mathcal{M}_{res}^{j}$ et un modèle de mouvement de caméra $\mathcal{M}_{cam}^{j}$ sont estimés à partir d'échantillons de vidéos formant la base d'apprentissage, pour chaque type j d'événement à détecter. Pour chaque segment vidéo s_i sélectionné précédemment, $z_i = (v_{res\,i}, \Delta v_{res\,i})$ représente les valeurs des mesures locales de mouvement et leurs contrastes temporels, et w_i représente les vecteurs de mouvement correspondant au modèle affine 2D estimé. Chaque segment s_i est ensuite étiqueté par un des J modèles d'événement dynamique appris, selon le critère du maximum de vraisemblance. Ainsi, l'étiquette l_i du segment s_i est définie comme suit :

$$l_i = \arg \max_{j=1,\ldots,J} P_{\mathcal{M}_{res}^{j}}(z_i) \times P_{\mathcal{M}_{cam}^{j}}(w_i) \tag{6}$$

5 Expérimentations

Dans cette partie, nous allons exposer les résultats obtenus sur un programme télévisé d'athlétisme. 10 minutes de cette séquence ont été consacrées à la phase d'apprentissage et 5 minutes à la phase de test. Le groupe "Play" est formé des segments de saut à la perche et de courses de fond, tandis que les segments "No play" contiennent des interviews et des vues d'ensemble du stade. A l'issue de la première étape de l'algorithme, comme le montre le premier tableau de la figure 1, seul un segment "No play" est étiqueté "Play", il correspond a une scène d'interview, mais il y a plus de mouvement en arrière-plan que dans les segments correspondants de la base d'apprentissage. Le taux de bonne détection pour le groupe "Play" est de 84% et le taux de fausse alarme est de 9%. Cette première étape donne donc des résultats satisfaisants. Le but est maintenant de détecter les événements pertinents de la vidéo d'athlétisme parmi les segments étiquetés "Play". Nous introduisons donc le modèle du mouvement de la caméra. Les quatre événements que nous voulons détecter sont les suivants : saut à la perche (SP), ralenti de saut à la perche (RP), plan large de la course de fond (LC), plan rapproché de la course de fond (RC). Le deuxième tableau de la figure 1

nous montre que la majorité des segments sont détectés de manière appropriée. Notons que le segment "No play" sélectionné lors de la première étape apparaît sur la ligne *NP*. Les erreurs de détection concernent deux segments appartenant à la classe "plan large de la course de fond". Elles sont dues au fait que le premier segment contient une scène de la course de fond que l'on peut situer entre un plan large et un plan rapproché et que le contenu du second segment concerné est presque similaire à une course d'élan du saut à la perche.

6 Conclusion

Nous avons présenté une approche originale et efficace pour la détection d'événements dans des vidéo, basée sur l'apprentissage de modèles probabilistes de mouvement. Elle tient compte explicitement des informations liées respectivement au mouvement de la scène et au mouvement de la caméra. Nous avons ainsi introduit un modèle probabiliste adapté des mesures locales du mouvement résiduel et de leurs contrastes temporels, ainsi qu'une modélisation appropriée du mouvement de la caméra. La méthode en deux étapes proposée pour la détection d'événements est générale et ne requiert pas de connaissances spécifiques en relation avec le genre de la vidéo, comme le type de sport considéré. Elle peut par conséquent s'appliquer à un large champ de vidéos. D'un autre côté, grâce au cadre statistique considéré, elle est assez flexible pour introduire proprement des a priori sur les classes si disponibles, ce qui conduirait à l'utilisation du critère MAP au lieu du critère MV, ou pour ajouter d'autres types d'informations tels que la couleur (la couleur dominante peut être utile, par exemple, pour tenir compte de la couleur du terrain ou du court de tennis dans les vidéos de sport) ou des descripteurs audio extraits de la bande son de la vidéo.

Références

1. C. Biernacki, G. Celeux, et G. Govaert. Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:719–725, 2000.
2. R. Fablet, P. Bouthemy, et P. Pérez. Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Trans. on Image Processing*, 11(4):393–407, 2002.
3. Y-F. Ma et H-J. Zhang. Motion pattern-based video classification retrieval. *EURASIP Journal on Applied Signal Processing*, 2:199–208, Mars 2003.
4. J. Nam et H. Tewfik. Dynamic video summarization and visualization. *7th ACM International Conference on Multimedia, ACM Multimedia'99*, Orlando, pages 53–56, Novembre 1999.
5. J-M. Odobez et P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Communication and Image Representation*, 6(4):348–365, Decembre 1995.
6. N. Peyrard et P. Bouthemy. Motion-based selection of relevant video segments for video summarisation. *Int. Conf. on Multimedia and Expo*, Baltimore, Juillet 2003.
7. Y. Rui et P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Hilton Head, SC, 1:111–118, 2000.
8. N. Vasconcelos et A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, Janvier 2000.
9. J. Vermaak, P. Pérez, et M. Gangnet. Rapid summarization and browsing of video sequences. 13^{th} *British Machine Vision Conference*, Cardiff, Septembre 2002.
10. Y. Yacoob et J. Black. Parametrized modeling and recognition of activities. *Sixth IEEE Int. Conf. on Computer Vision*, Bombay, India, pages 120–127, 1998.
11. L. Zelnik-Manor et M. Irani. Event-based video analysis. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2:123–130, Decembre 2001.