

Détection supervisée d'événements à l'aide d'une modélisation probabiliste du mouvement perçu

Supervised event detection using probabilistic models of visual motion

G. Piriou¹

F. Coldefy¹

P. Bouthemy¹

J.-F. Yao^{1,2}

¹ IRISA/INRIA, Projet Vista

² IRMAR

Campus universitaire de Beaulieu, 35042 Rennes cedex, France.

Gwenaelle.Piriou@irisa.fr

Résumé

Cet article présente une méthode supervisée de détection d'événements dans une vidéo, à partir de l'analyse de son contenu dynamique. Par souci de généralité et d'efficacité, l'information de mouvement est captée par des mesures simples. Elle est de plus décomposée en mouvement dominant et mouvement résiduel dans l'image. Ces mesures sont représentées par des modèles probabilistes parcimonieux permettant de placer le problème traité dans un cadre bayésien. Pour induire une phase d'apprentissage faiblement supervisée et augmenter la robustesse de la détection, notre méthode s'articule en deux étapes principales : sélection des segments vidéos susceptibles d'appartenir au monde des événements pertinents, détection des événements d'intérêt parmi les segments vidéos retenus. La mise en oeuvre de cette méthode nécessitant une segmentation préalable de la vidéo en plages homogènes, nous décrivons une technique de segmentation temporelle basée sur l'analyse du mouvement dominant dans l'image. Des résultats expérimentaux sur des vidéos de sport sont présentés.

Mots Clef

Mesures de mouvement, modélisation probabiliste du mouvement, classification supervisée d'événements, segmentation de vidéos.

Abstract

This paper presents a supervised method for detecting relevant events in videos, based on the analysis of their dynamic content. For generality and efficiency purposes, motion information is captured from simple motion-related measurements. Moreover, it is decomposed into dominant and residual image motion. The measurements are then represented by probabilistic models which allow us to formulate the detection problem in a bayesian framework. To infer a weakly supervised learning stage and to increase the robustness of the event detection, our method proceeds in two

steps. The first one selects video segments which are supposed to belong to the world of relevant events. The second one detects the events of interest among the selected segments. Since the video to be analysed first needs to be segmented into homogeneous temporal units, we also describe a segmentation technique exploiting the dominant image motion. Experimental results on sport videos are reported.

Keywords

Motion measurements, probabilistic motion modelling, supervised event classification, video segmentation.

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 concerne notamment des domaines comme l'indexation vidéo ou la surveillance de scènes. La principale difficulté réside dans la détection d'événements de nature "sémantique" à partir d'informations de bas niveau extraites des images. Les caractéristiques d'un événement visuel doivent ainsi être exprimées en termes de primitives vidéo (couleur, texture, mouvement, forme ...) suffisamment discriminantes relativement à son contenu. Plusieurs approches ont été développées exploitant différentes sortes de primitives vidéo. Les auteurs de [14] se sont intéressés à la classification de vidéos en différents genres (sport, journaux télévisés, films, publicités, documentaires, ...) en introduisant des modèles statistiques sur des éléments structurant une vidéo, à savoir, la durée des plans et leur activité. Dans [8], une méthode de classification basée sur les SVM ("support vector machines"), utilisant un descripteur de "texture de mouvement" a été présentée. L'application visée dans [10] est l'indexation de vidéos en utilisant une méthode de clustering combinant information de mouvement et couleur. La combinaison d'informations extraites des bandes image et audio est également exploitée, par exemple dans [9], pour la création de résumé

de vidéo.

Comme le contenu de nature dynamique d'une vidéo est un indicateur important de l'apparition ainsi que du type d'un événement, l'analyse du mouvement dans les séquences d'images est évidemment une voie d'investigation privilégiée pour la segmentation des vidéos en plages significatives et pour la reconnaissance d'événements particuliers. Une caractérisation du mouvement peut être dérivée des champs denses de vitesses, comme dans [13] pour la détection d'activités domestiques. Dans [16], les auteurs utilisent des mesures spatio-temporelles très simples, à savoir des histogrammes des dérivées spatiales et temporelles de l'intensité, pour repérer des événements particuliers. Dans [15], 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. Les auteurs de [4] modélisent quant à eux le mouvement sur une base d'ondelettes et appliquent cette modélisation à l'indexation de séquences vidéos.

Nous nous intéressons au problème de la détection d'événements d'intérêt à partir de l'information de mouvement. L'objectif général que nous poursuivons est de concevoir une méthode générique, mais flexible (c.a.d spécialisable), s'appuyant sur des formulations statistiques, la sémantique étant apportée par une phase d'apprentissage supervisé. Nous avons déjà développé une première approche dans [5, 6, 12]. Le travail décrit dans cet article l'étend à plusieurs titres. Nous cherchons en effet à en accroître l'économie de calcul (primitives de mouvement considérées), la parcimonie de la modélisation, et la puissance de représentation (séparation des composantes de mouvement et apprentissage des classes d'événements). La méthode proposée fonctionne de plus en deux étapes principales. La première étape est une phase de tri des segments vidéos extraits de la vidéo traitée, entre ceux relevant du "monde des événements" que l'on cherche à détecter et les autres. La seconde étape est alors consacrée à la reconnaissance des événements pertinents au sein des segments retenus. En procédant de la sorte, nous visons une modélisation ciblée et appropriée des contenus, un traitement allégé, et au final, une plus grande robustesse.

Dans la section 2, nous définissons les modèles probabilistes de mouvement nécessaires à la méthode de détection d'événements qui sera décrite à la section 3. Une validation comparative du modèle de mouvement résiduel proposé est fournie en section 4. Puis, nous décrivons dans la section 5 la façon dont est réalisée l'étape préalable de segmentation de la vidéo, avant de donner des résultats expérimentaux de reconnaissance d'événements en section 6. La conclusion formera la section 7.

2 Modèles probabilistes de mouvement

Notre méthode de détection implique tout d'abord une modélisation probabiliste du mouvement contenu dans la vidéo.

En effet, la variabilité des contenus et l'absence de modélisation analytique rendant compte de tous les mouvements possibles, conduit naturellement à privilégier une approche probabiliste, qui de plus permettra de poser le problème de reconnaissance des événements dans un cadre bayésien. Par souci de généralité et d'efficacité, seules des mesures locales de mouvement de bas niveau sont exploitées. Ces mesures sont calculées directement à partir des dérivées spatio-temporelles de l'intensité lumineuse. L'information relative au contenu de la scène et celle apportée par le mouvement de la caméra étant complémentaires, il nous faut en fait introduire un modèle probabiliste pour chacune de ces composantes de mouvement.

2.1 Modélisation probabiliste du mouvement de la scène

Notre première approche de ce problème avait été dans [5, 6] de caractériser directement le mouvement perçu dans la vidéo à partir de quantités locales correspondant aux moyennes pondérées d'amplitudes des 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 résultante. 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. Le mouvement dominant sera souvent associé au mouvement de la caméra et le mouvement résiduel au mouvement de la scène dans le descriptif qui va suivre pour illustrer plus aisément notre propos, mais ceci n'est en fait pas obligatoire.

A chaque instant t , un modèle de mouvement 2D paramétrique (noté par le vecteur de paramètres θ_t) représentant le mouvement dominant dans l'image est estimé comme expliqué au paragraphe 2.2. Les mesures locales de mouvement résiduel $v_{res}(p, t)$ sont alors définies comme la moyenne pondérée par le carré de la norme du gradient spatial d'intensité, des amplitudes de vitesses normales résiduelles v_n . Ces dernières sont en fait déduites de la différence d'image déplacée fournie par le mouvement dominant estimé et notée $DFD_{\hat{\theta}_t}$, en effet, $v_n(p, t) = \frac{|DFD_{\hat{\theta}_t}(p)|}{\|\nabla I(p, t)\|}$. On obtient après simplification :

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

où $DFD_{\hat{\theta}_t}(q) = I(q + \mathbf{w}_{\hat{\theta}_t}(q), t + 1) - I(q, t)$ avec $\mathbf{w}_{\hat{\theta}_t}(q)$ le vecteur de vitesse fourni au pixel q par le modèle de mouvement estimé. $\mathcal{F}(p)$ est une fenêtre spatiale au point p et $N_{\mathcal{F}(p)}$ représente le nombre de pixels qu'elle contient. $\nabla I(q, t)$ 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.

Dans [5, 6, 12], les mesures locales de mouvement étaient ensuite quantifiées, puis nous considérons la distribution de leurs co-occurrences temporelles. Cette distribution était ensuite modélisée par un modèle causal de Gibbs [5, 6]. Cependant, un tel modèle comporte un très grand nombre de paramètres. Pour réduire la taille des modèles mis en jeu, nous avons proposé dans [12] de considérer la matrice des co-occurrences temporelles comme un histogramme empirique 2D et de la représenter par un mélange gaussien 2D. Nous présentons ici un modèle encore plus parcimonieux tout en étant encore porteur de suffisamment d'informations pour un objectif de détection d'événements particuliers dans une vidéo. Nous considérons les mesures locales de mouvement (1) telles quelles. Pour différentes séquences vidéos, les histogrammes de ces quantités cumulées sur le temps se sont avérés être relativement proches d'une distribution gaussienne (tronquée, par définition, sur les valeurs positives), complétée d'un pic en zéro. La distribution des mesures locales de mouvement est, par conséquent, approchée par un modèle de mélange spécifique de densité $f_{v_{res}}$ relativement à la mesure $\mu(dx) = \delta_0(dx) + dx$, où δ_0 représente la mesure de Dirac en 0. Nous avons :

$$f_{v_{res}}(x) = \beta\delta_0(x) + (1 - \beta)\phi_t(x; 0, \sigma^2) \quad (2)$$

β est le poids du mélange, δ_0 est la fonction de Dirac ($\delta_0(x) = 1$ si $x = 0$ et $\delta_0(x) = 0$ sinon) et ϕ_t est la densité d'une distribution gaussienne tronquée, 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 tout de même l'évolution temporelle de l'information de mouvement, nous considérons les contrastes temporels Δv_{res} des mesures locales de mouvement. Ces contrastes sont définis en chaque pixel p de l'image et à chaque instant t de la façon suivante :

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

La distribution des Δv_{res} est à nouveau modélisée par un mélange d'une gaussienne centrée, mais cette fois non tronquée, et d'une mesure de Dirac en 0. Les deux paramètres de ce modèle (le poids du mélange et la variance de la distribution gaussienne) sont également déterminés par MV. Finalement, le modèle global du mouvement résiduel est défini comme le produit des deux modèles introduits :

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

Ce modèle a l'avantage d'être complètement spécifié par quatre paramètres, qui s'estiment très facilement. Le temps de calcul est donc considérablement réduit par rapport aux deux modèles déjà mentionnés [5, 6, 12]. Naturellement, du fait de sa grande simplicité, il possède aussi ses limites. En effet, en traitant toutes les mesures "en vrac" à travers leurs histogrammes cumulés sur tous les points et les images du segment vidéo considéré, nous ne captions pas leurs dépendances spatiales et temporelles. Pour l'application visée, ce manque de spatialisation n'a pas d'impact

fort, car les événements que nous souhaitons détecter ne sont pas liés à une localisation particulière dans l'image. Quant à l'aspect temporel, il est capté en partie, à travers les contrastes temporels des mesures locales de mouvement. Néanmoins, comme nous n'avons aucune information sur l'historique (que pourrait apporter par exemple le calcul de trajectoires de points), nous chercherons à détecter des catégories d'événements et non des situations très précises. Par ailleurs, il suffit que les modèles construits soient suffisamment discriminants pour réaliser les deux phases (tri puis reconnaissance) de la méthode de détection d'événement, spécifiée ainsi justement à dessein. Ces modèles n'ont pas à être complètement explicatifs des contenus d'autant que nous ne considérons qu'un "univers" limité et appris d'événements.

2.2 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 de la phase de reconnaissance. Premièrement, le mouvement dominant dans l'image est représenté par un modèle affine 2D déterministe :

$$\mathbf{w}_\theta(p) = \begin{pmatrix} a_1 + a_2x + a_3y \\ a_4 + a_5x + a_6y \end{pmatrix}, \quad (5)$$

où $\theta = (a_i, i = 1, \dots, 6)$ est le vecteur de paramètres du modèle et p est un point de l'image dont les coordonnées (x, y) sont exprimées dans le repère d'origine le centre 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 [11].

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 [12]. 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 appréhendable assez aisément, la tâche se complique pour ce qui est des paramètres du premier ordre qui ne sont pas constants sur un segment et qui ne sont pas non plus de même nature. Pour cette raison, nous proposons de construire la carte des vecteurs de vitesses associés au mouvement dominant, 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 $\mathbf{w}_{\hat{\theta}_t}(p)$ sont calculés en chaque point p du support de l'image. Les composantes horizontales et verticales des $\mathbf{w}_{\hat{\theta}_t}(p)$ sont ensuite finement quantifiées afin de construire l'histogramme 2D empirique de leur distribution sur le segment vidéo. Finalement, cet histogramme est représenté par un mélange de distributions gaussiennes. Le nombre de composantes

dans le mélange est déterminé par le critère ICL (Integrated Completed Likelihood, [2]) et les paramètres du modèle sont ensuite estimés par l’algorithme EM.

3 Méthode de détection d’événements

Les modèles probabilistes de mouvements introduits vont être exploités pour la détection d’événements dans une vidéo. La vidéo considérée est tout d’abord segmentée en plages temporelles homogènes par une méthode qui sera exposée à la section 6. Nous noterons $\{s_i\}_{i=1,\dots,N}$ la partition de la vidéo en plages homogènes. La méthode de détection proposée se décompose en deux grandes étapes. Dans un premier temps, l’ensemble $\{s_i\}_{i=1,\dots,N}$ des segments vidéos est réduit de façon à ne contenir que les segments susceptibles de représenter les événements cherchés. Il est évident que pour chaque genre de vidéo, il faudra définir ce qu’est censé contenir le “monde des événements”. Lors de la deuxième étape, chaque segment retenu sera étiqueté par un des événements considérés.

3.1 Tri initial des segments vidéos

Pour ne pas complexifier inutilement les modèles dans cette phase initiale et alléger d’autant l’apprentissage associé, nous ne considérons pour ce tri, que le mouvement résiduel généralement porteur du mouvement de la scène. Pour un genre donné de vidéo, nous procédons à une phase d’apprentissage au cours de laquelle plusieurs modèles de mouvement sont appris, d’une part, pour le monde des événements, et d’autre part, pour l’ensemble des segments n’appartenant pas au monde des événements. Plus précisément, sur un ensemble de vidéos servant d’exemples, sur lequel la vérité-terrain (séparation en les deux groupes spécifiés plus haut) est établie, nous effectuons une classification hiérarchique ascendante des modèles de mouvement résiduel sur les segments du monde des événements, puis sur l’ensemble des segments de l’autre groupe. L’objectif est de représenter chacun de ces deux groupes par quelques “clusters” appropriés, nécessaires pour rendre compte de l’hétérogénéité de chacun de ces deux grands groupes de segments. Le déroulement de cette classification hiérarchique ascendante est décrit ci-dessous :

- *Pas 0* : Un modèle de mouvement résiduel \mathcal{M}_{res} , de la forme (4), est estimé pour chaque segment du groupe considéré. Chaque segment forme un cluster distinct au départ.
- *Pas 1* : On réunit les deux clusters dont les lois sont les plus proches au sens de la distance de Kullback-Leibler symétrisée (la formule est donnée en annexe). On estime un nouveau modèle pour le cluster ainsi formé. Tant que le *critère d’arrêt* n’est pas vérifié, on itère.
- *Critère d’arrêt* : Il consiste à comparer le minimum des distances de Kullback-Leibler symétrisées entre clusters (formés lors de l’étape considérée) au minimum des distances entre clusters au pas 0.

À l’issue de cet apprentissage automatique, nous disposons d’un ou plusieurs modèles de mouvement pour chacun des groupes. On notera N_1 (respectivement N_2) le nombre de modèles retenus pour le monde des événements (respectivement, pour le reste). L’atout majeur de cette méthode de classification hiérarchique ascendante est d’être simple et automatique, et surtout d’éviter de construire “à la main” des classes sémantiques sur l’ensemble des contenus possibles des vidéos considérées.

Le tri consiste à attribuer une étiquette binaire ζ_i à chaque segment s_i de la vidéo traitée, selon le critère du maximum de vraisemblance :

$$\zeta_i = \arg \max_{k=1,2} \max_{1 \leq n \leq N_k} P_{\mathcal{M}_{res}^{k,n}}(z_i) \quad (6)$$

$z_i = (v_{res\ i}, \Delta v_{res\ i})$ représente les valeurs des mesures locales de mouvement et leurs contrastes temporels pour le segment s_i .

3.2 Détection des événements pertinents

Après le tri initial, nous sommes supposés travailler sur “l’espace fermé” qu’est le monde des événements, notre problème de détection d’événements pouvant alors se transformer en un problème de reconnaissance. Ce dernier est plus aisé à traiter, car il s’agit de retrouver la meilleure classe parmi un nombre prédéfini, pour un ensemble réduit de segments vidéos.

À ce stade, les deux types d’information (mouvement résiduel et mouvement de caméra) sont requis, puisque leur combinaison permet de caractériser plus précisément un événement particulier. Lors de la phase d’apprentissage supervisé associé à cette étape de reconnaissance, un modèle de mouvement résiduel \mathcal{M}_{res}^j et un modèle de mouvement de caméra \mathcal{M}_{cam}^j sont estimés à partir des segments de la base d’apprentissage, pour chaque type j d’événement à détecter.

Pour chaque segment vidéo s_i retenu à l’issue de l’étape de tri, $z_i = (v_{res\ i}, \Delta v_{res\ i})$ représente les valeurs des mesures locales de mouvement résiduel et leurs contrastes temporels, et w_i représente les vecteurs de mouvement correspondant au mouvement dominant estimé. Chaque segment s_i est cette fois étiqueté selon le critère du maximum de vraisemblance par un des J événements représentés par les modèles de mouvement combinés appris. Ainsi, l’étiquette l_i du segment s_i est définie comme suit :

$$l_i = \arg \max_{j=1,\dots,J} P_{\mathcal{M}_{res}^j}(z_i) \times P_{\mathcal{M}_{cam}^j}(w_i) \quad (7)$$

Si disponible ou établi par ailleurs, un a priori sur chaque classe pourrait être ajouté au critère (7).

4 Validation du modèle de mouvement résiduel

Afin de valider le modèle probabiliste du mouvement résiduel décrit au paragraphe 2.1, que nous notons MDG, nous le



FIG. 1 – Danse sur glace : images représentatives de la séquence.

comparons avec le modèle de [12] consistant à représenter les co-occurrences temporelles par un modèle de mélange 2D gaussien, noté MG2D. Cette comparaison est effectuée sur la phase de tri des segments. En fait, comme il est difficile d'estimer la distance de Kullback-Leibler entre deux modèles de mélange 2D gaussien, la méthode de classification hiérarchique ascendante n'est pas utilisée. Nous considérons donc un cadre entièrement supervisé. Il y a autant de modèles appris que de classes sémantiques prédéfinies pour chacun des deux groupes considérés pour le tri. Du fait de la limitation en nombre de pages, nous exposons seulement les résultats obtenus sur une vidéo. Il s'agit d'un extrait de programme télévisé de danse sur glace. La figure 1 montre des images représentatives de cet extrait. 23 minutes de cette séquence vidéo ont été utilisées pour la phase d'apprentissage et 9 minutes pour la phase de test. Pour bien appréhender l'impact de la modélisation sur la détection des événements pertinents, formés en l'occurrence par les figures artistiques, les ralentis et les glissades des patineurs, nous partons d'une segmentation manuelle de la vidéo. Trois modèles de mouvement sont donc appris pour ces trois classes. Quatre modèles de mouvement représentent les quatre classes sémantiques qui forment le reste des segments, à savoir : vues du public, scènes statiques (début et fin de la représentation), salut des patineurs et attente des notes par les patineurs. 41 segments vidéos de la base test (sur 63 au total) représentent des événements pertinents. Le tableau 1 contient les résultats du tri des segments vidéos pour les deux types de modélisation considérés. Le taux de précision, noté P , et le taux de rappel, noté R , sont définis

de la façon suivante :

$$P = \frac{\#\text{corrects}}{\#\text{corrects} + \#\text{intrus}}$$

$$R = \frac{\#\text{corrects}}{\#\text{corrects} + \#\text{manqués}}$$

où $\#\text{corrects}$ est le nombre de segments étiquetés comme appartenant au monde des événements et représentant effectivement un des événements, $\#\text{intrus}$ est le nombre d'intrus (segments étiquetés de manière erronée) dans le monde des événements et $\#\text{manqués}$ est le nombre de segments représentant un événement mais n'ayant pas été étiquetés comme tels.

Alors que la modélisation exploitant les co-occurrences est supposée capter plus d'informations, il apparaît que la modélisation introduite dans cet article utilisant les mesures locales de mouvement et leurs contrastes temporels, donnent des résultats de même nature, tout en conduisant à un temps de calcul plus faible. En effet, pour le modèle MG2D, le temps de calcul est de 0.8 sec/image avec un Pentium IV 2.4 Ghz, tandis qu'il est de 0.2 sec/image pour le modèle MDG.

	P	R
MG2D	1	0.83
MDG	0.95	0.90

TAB. 1 – Danse sur glace : Résultats de l'étape de tri des segments pour les modélisations MG2D et MDG. P = taux de précision, R = taux de rappel.

5 Segmentation temporelle de la vidéo

5.1 Choix des descripteurs

Avant de procéder à la détection d'événements dans une vidéo, nous devons disposer d'une segmentation temporelle de cette vidéo telle que chaque segment obtenu ne soit supposé contenir qu'un type d'activité pour que la phase de reconnaissance spécifiée plus haut puisse s'appliquer. Un découpage en plans de la vidéo n'est donc pas suffisant. La segmentation doit s'appuyer sur l'information de mouvement. Pour des raisons d'efficacité, nous n'exploitons que le mouvement dominant représenté par les paramètres $(a_i)_{i=1..6}$ du modèle affine (5). En fait, il est plus simple de considérer un jeu de vecteur $w_\theta(p)$ issu de ce modèle pour quelques positions particulières de points p , et de spécifier des descripteurs appropriés à partir de ces vecteurs. Plusieurs alternatives peuvent être considérées. S'il suffit de repérer les changements de direction des travellings et des panoramiques, l'étude de l'évolution dans le temps du seul vecteur $w_\theta(0,0) = (a_1, a_4)$ au centre de l'image permet de détecter la présence de travellings ou de panoramiques ainsi que leur changement de direction. En revanche, l'observation de ce seul vecteur ne permet pas de déceler systématiquement la présence de zooms qui influent sur les paramètres a_2 et a_6 . Si les zooms doivent être

précisément pris en compte, on s'intéressera au vecteur de déplacement calculé en un point (ou plusieurs points) où les 6 paramètres du modèle affine interviennent (en un des coins de l'image par exemple). La norme de ce vecteur permet de distinguer un mouvement dominant nul de la présence de panoramiques, de travellings ou de zooms. La variation de son angle caractérise un changement d'orientation du mouvement dominant. Plus généralement, il suffit d'étudier les vecteurs de déplacement calculés en trois endroits bien choisis de l'image (au centre et aux coins supérieurs gauche et droit de l'image par exemple) pour rendre compte de la variabilité des six paramètres du modèle affine. Le choix de ces vecteurs dépend du niveau de granularité de la segmentation temporelle qu'on souhaite obtenir. Dans le cadre de notre application impliquant des vidéos de sport, nous nous focalisons sur la détection des changements de direction du mouvement dominant qui traduisent une évolution du contenu de la scène, puisque, en général, la caméra suit les joueurs ou les athlètes. Dans ce but, nous nous intéressons à la norme et à l'orientation du vecteur de mouvement $w_\theta(0,0)$ pris au centre de l'image.

5.2 Définition du critère de segmentation

En plus des mouvements de caméra, il faut aussi détecter les ruptures de montage (cuts, fondus, etc.) qui seront appréhendées par l'analyse de la zone portant le mouvement dominant dans l'image, ou en d'autres termes, pour laquelle le mouvement dominant estimé explique la totalité du mouvement observé, comme proposé dans [3]. On s'intéresse en fait à la taille normalisée ζ_t de cette zone. Cette quantité est à peu près constante au sein d'un même plan, mais subit une forte variation lors d'une rupture de plan.

Comme indiqué plus haut, les variations du mouvement dominant au sein d'un plan sont identifiées par l'étude de l'évolution moyenne de l'amplitude ν_t et de l'orientation β_t du vecteur $w_\theta(0,0) = (a_1, a_4)$. La segmentation en plages d'activité homogène est obtenue en combinant la détection des variations du support ζ_t , de l'angle β_t , et de l'amplitude ν_t .

Les ruptures de montage se manifestent sur le support ζ_t par des variations de moyenne ou par des chutes locales de cette quantité. La détection des sauts de moyenne est réalisée en appliquant le test cumulatif de Page-Hinkley ([1]). Ce test est connu pour sa robustesse et son faible coût en temps de calcul. La détection des chutes locales de ζ_t est obtenue en utilisant un test gaussien sur les valeurs aberrantes [7]. En pratique, ces deux tests sont mis en place en parallèle. Le test de Hinkley est défini par les détecteurs de sauts décroissants ou croissants suivants:

$$M_0 = 0, \quad M_n = \sum_{t=1}^n (\zeta_t - \mu_0 + \frac{j_{min}}{2})$$

$$m_n = \max_{0 \leq t \leq n} M_t, \text{ alarme si } m_n - M_n > \lambda,$$

$$U_0 = 0, \quad U_n = \sum_{t=1}^n (\zeta_t - \mu_0 - \frac{j_{min}}{2})$$

$$u_n = \min_{0 \leq t \leq n} U_t, \text{ alarme si } U_n - u_n > \lambda,$$

où j_{min} peut s'interpréter comme l'écart toléré a priori et λ

un seuil prédéfini. La moyenne inconnue μ_0 est estimée en ligne et réinitialisée après chaque saut. L'instant de rupture est le dernier indice n pour lequel $m_n = M_n$ ou $u_n = U_n$. Le test sur les valeurs aberrantes est appliqué en chaque instant t selon:

$$\text{alarme si } |\zeta_t - \mu_0|/\sigma_0 > s$$

où s est un seuil gaussien prédéfini, et σ_0 l'écart-type inconnu, calculé en ligne. s est pris égal à 3.

Les coordonnées cartésiennes du vecteur de déplacement $w_\theta(0,0)$ peuvent être bruitées (par exemple pour des plans rapprochés d'objets en mouvement) et sont lissées par un filtre médian temporel de taille 9. La norme ν_t et l'angle β_t sont calculés après ce filtrage. La fonction β_t est la fonction d'angle de \mathbb{R}^2 dans \mathbb{R} assurant $|\beta_t - \beta_{t+1}| < \pi$ et par convention $-\pi < \beta_0 \leq \pi$.

Les variations d'orientation sont détectées par un test de Hinkley pour une valeur de saut j_{min} dépendant de l'application. Typiquement, si on souhaite détecter des changements de direction latéraux ou verticaux on prendra une valeur de saut proche de π . Pour plus de sensibilité, on choisira des valeurs entre $\pi/4$ et $\pi/2$.

La segmentation de ν_t est obtenue par simple seuillage $\nu_t \leq c$ et $\nu_t > c$ où c est une valeur prédéfinie. La segmentation est régularisée afin de supprimer les points isolés et de favoriser l'émergence de segments plutôt longs.

La fusion des détections de rupture de montage et des changements de direction et d'amplitude du déplacement de la caméra est réalisée de façon hiérarchique. Les détections sur le support ζ_t sont prises en compte intégralement sans remise en cause. La segmentation sur l'amplitude ν_t est ensuite ajoutée, recalée avec les détections de rupture sur ζ_t ; il arrive en effet qu'une rupture de montage soit détectée à la fois sur le support et l'amplitude, avec un léger décalage de trames (une ou deux). Les segments trop courts (moins de 10 trames) sont supprimés. Enfin, on intègre les détections de changement de direction. Seules sont prises en compte les détections réalisées sur les plages de mouvement dominant significatif ($\nu_t > c$).

5.3 Résultats de segmentation temporelle

Nous donnons ci-après deux exemples de segmentation automatique pour des séquences d'athlétisme. Les paramètres du test de Hinkley sur le support sont $j_{min} = 0.2$, et $\lambda = 2$. Les valeurs de j_{min} et λ ont été fixées empiriquement à partir de l'évaluation de la segmentation temporelle réalisée sur 2 heures de vidéos. Le seuil c sur l'amplitude est pris égal à 2. Enfin, les paramètres du test de Hinkley sur l'angle β_t sont $j_{min} = \pi/2$ et $\lambda = 2$.

La figure 2 présente la segmentation automatique obtenue pour une séquence de saut à la perche. Chaque image représente l'image médiane des segments. La première plage est un plan fixe pendant lequel l'athlète s'élanche. Elle est suivie par un zoom arrière au fur et à mesure que le perchiste s'approche. Le saut est décomposé en deux segments, la montée puis la retombée. La dernière plage correspond au

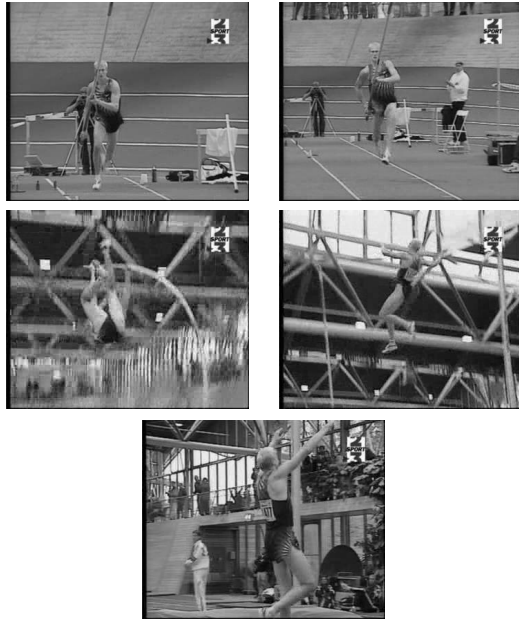


FIG. 2 – Segmentation temporelle d’une séquence de saut à la perche : images médianes des segments obtenus (vidéo fournie par l’INA).

suivi du perchiste en train de se relever. La figure 3 détaille la vérité terrain du découpage au sens du mouvement dominant (bandeau a), la segmentation obtenue (bandeau b) ainsi que les détections de ruptures sur le support ζ_t (bandeaux c et d), et les segmentations de l’amplitude ν_t (bandeaux e et f) et enfin les détections de variations de l’angle β_t (bandeaux g et h). Globalement, vérité terrain et segmentation automatique s’accordent plutôt bien, à un léger décalage temporel près. Si deux détections sont très proches l’une de l’autre, elles sont fusionnées.

La figure 4 présente les images médianes des plages obtenues pour la segmentation automatique d’une séquence de saut en longueur. Les bandeaux de la vérité terrain et de la segmentation temporelle sont présentés à la figure 5. Là encore, segmentation automatique et réalité terrain s’accordent bien. La première plage correspond à la phase d’élan qui est un panoramique latéral. Le saut est contenu dans la seconde plage qui est associée à un zoom arrière. La troisième plage correspond à un plan fixe sur l’athlète se relevant et enfin la dernière plage est un suivi de l’athlète. Les résultats obtenus montrent une segmentation plutôt en accord avec les principales variations du mouvement dominant et soulignent l’intérêt de combiner les trois descripteurs retenus associés au mouvement dominant entre images. La localisation temporelle des changements de mouvement est parfois encore imprécise et peut être sans doute améliorée. La segmentation temporelle obtenue n’est pas à l’abri de sur-segmentations ou de sous-segmentations locales. Les choix des différents paramètres de l’algorithme sont le fruit d’un compromis sur les tests réalisés sur environ 2 heures de vidéos.

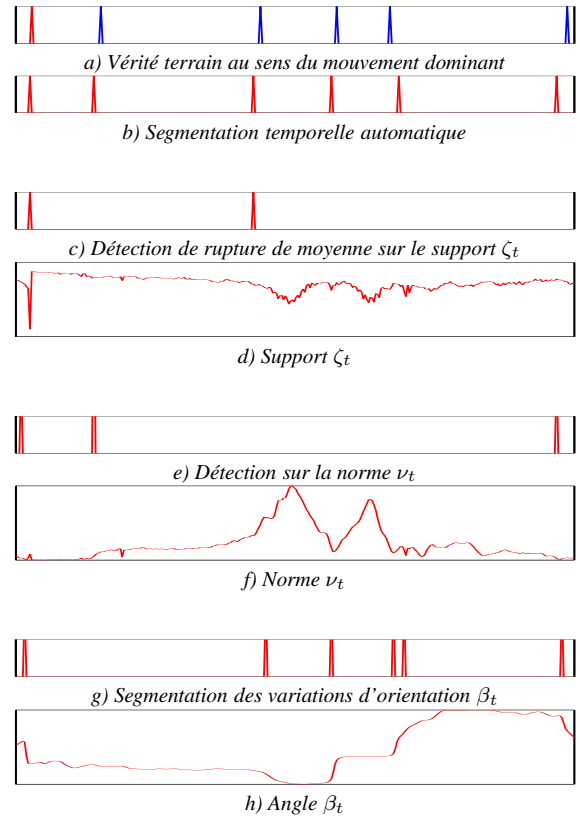


FIG. 3 – Segmentation temporelle d’une séquence de saut à la perche : découpage obtenu à partir des trois descripteurs considérés.

6 Evaluations expérimentales de la méthode de détection d’événements

Nous présentons dans cette section les résultats de notre méthode de détection pour deux séquences vidéos de sport. Le choix de telles séquences n’est pas anodin. En effet, dans ce type de vidéo, l’aspect sémantique est très corrélé avec le mouvement, à la fois le mouvement dans la scène et le mouvement de la caméra, ce qui correspond tout à fait au contexte que nous avons étudié. Le monde des événements sera formé des phases de jeux ou d’action (segments contenant du mouvement) et le reste des segments représenteront les “temps morts”. De plus, les diffusions de temps forts de manifestations sportives étant très fréquentes, le fait de détecter les événements pertinents dans une vidéo d’un sport donné trouve des applications immédiates et correspond à une véritable demande dans le domaine du multimédia professionnel et grand public.

6.1 Séquence vidéo d’athlétisme

Nous avons tout d’abord traité un extrait de programme télévisé d’athlétisme. 10 minutes de cette séquence vidéo ont été utilisées pour la phase d’apprentissage et 6 minutes pour la phase de test. Comme illustré sur la figure 6, l’extrait considéré est formé de sauts à la perche et de course

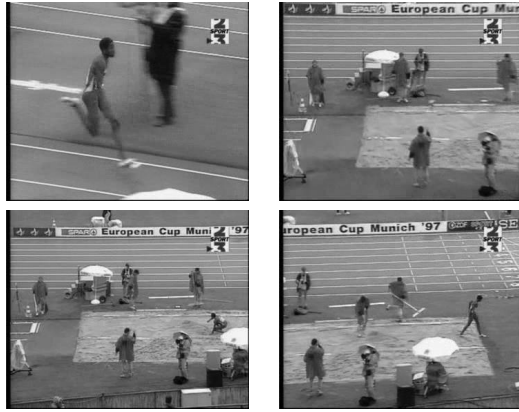


FIG. 4 – Segmentation temporelle d’une séquence de saut en longueur : images médianes des segments obtenus (vidéo fournie par l’INA).

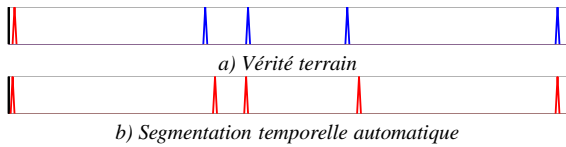


FIG. 5 – Segmentation temporelle d’une séquence de saut en longueur : vérité terrain et découpage final obtenu.

de fond, entrecoupés d’interview et de vues du stade.

Tri initial des segments vidéos.

Les événements considérés sont le saut à la perche et la course de fond. L’algorithme de classification hiérarchique ascendante appliqué sur l’ensemble des segments de la base d’apprentissage par deux fois (une fois sur le monde des événements et l’autre fois sur le reste des segments) retient 10 modèles de mouvement résiduel pour le monde des événements et 2 pour le reste des segments. Nous obtenons les résultats suivants pour le tri : le taux de précision est de 0.98 et le taux de rappel de 0.93 . Autrement dit, 98% des segments étiquetés comme appartenant au monde des événements sont effectivement des segments de saut à la perche ou de course de fond, et 93% des segments de saut à la perche et de course de fond sont étiquetés comme tels. Il y a seulement un segment “intrus” dans le monde des événements. Ce segment représente une scène d’interview, avec parfois des passages de personnes devant le journaliste et l’athlète, ainsi que du passage en arrière-plan. À ce stade, les résultats sont donc plutôt satisfaisants.

Détection d’événements.

Nous nous intéressons maintenant à la phase de reconnaissance des événements pertinents qui combine le mouvement résiduel (lié à la scène) et le mouvement dominant (lié à la caméra). Les quatre événements à détecter parmi les segments retenus sont les sauts à la perche (Perche), les ralentis de saut à la perche (Perche-R), les plans larges



FIG. 6 – Athlétisme : De gauche à droite et de haut en bas : saut à la perche, plan large de course de fond, plan rapproché de course de fond, interview, et vue d’ensemble du stade (vidéo fournie par l’INA).

de courses de fond (Course-PL) et les plans rapprochés de course de fond (Course-PR). Les résultats sont répertoriés dans le tableau 2. La première colonne de ce tableau désigne la vérité terrain. Nous retrouvons naturellement dans ce tableau le segment intrus sélectionné lors de la phase de tri. Les erreurs de classification concernent essentiellement des plans larges de courses de fond, étiquetés comme étant des ralentis de saut à la perche puisque le ralenti de saut à la perche inclut la course d’élan. Par ailleurs, les résultats sont probants. La figure 7 montre des images représentatives de segments bien étiquetés pour chaque événement à détecter.

	Etiquette attribuée			
	Perche	Perche-R	Course-PL	Course-PR
Perche	4	0	0	0
Perche-R	0	10	0	1
Course-PL	0	7	14	3
Course-PR	0	0	1	17
Intrus	0	1	0	0

TAB. 2 – Athlétisme : Résultats de l’étape de reconnaissance (en nombre de segments).

6.2 Séquence vidéo de tennis

Cette deuxième séquence est un extrait d’un match de tennis. La base d’apprentissage est formée de 22 minutes et nous avons testé notre méthode sur les 15 minutes suivantes. Le monde des événements regroupera les échanges et les services. Le reste des segments représente des vues

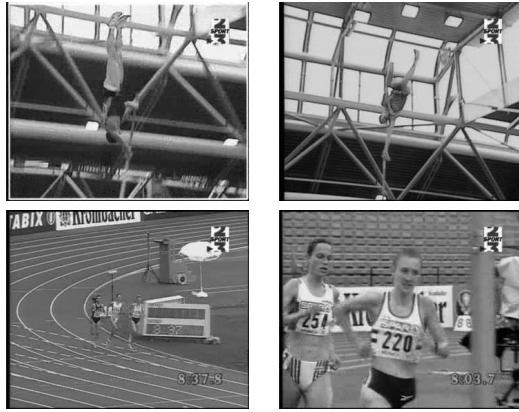


FIG. 7 – Athlétisme : Images représentatives de segments bien étiquetés lors de la phase de détection (saut à la perche, ralenti de saut, plan large de course de fond, plan rapproché de course de fond).

du public, les joueuses qui se replacent après un échange, etc ... Des images extraites de cette séquence apparaissent à la figure 8.

Tri initial des segments vidéos.

Dans la base d'apprentissage, nous avons 86 segments appartenant au monde des événements et 149 segments extérieurs à cet ensemble. L'algorithme de classification hiérarchique ascendante retient 14 modèles résiduels pour le monde des événements et 42 modèles pour le deuxième groupe. Le nombre élevé de modèles retenus, surtout pour l'ensemble des segments n'appartenant pas au monde des événements, montre la diversité des contenus dynamiques de cette vidéo. Les résultats du tri des segments vidéos figurent dans le tableau 3. Le taux de précision est ici de 0.79 et le taux de rappel vaut 0.76.

	Etiquette attribuée	
	ME	Reste
ME	59	19
Reste	16	110

TAB. 3 – Tennis : Tri initial des segments vidéos (en nombre de segments). ME = Monde des événements, Reste = Ensemble des segments étrangers au monde des événements.

Détection d'événements.

Les événements que nous cherchons à détecter sont les échanges et les services. Notons que les services peuvent être filmés de deux façons : soit la joueuse est filmée au niveau du buste, soit elle est filmée en entier. On considère donc trois événements. Le tableau 4 donne la matrice de confusion issue de l'étape de reconnaissance pour ces trois événements.

Ces résultats tout à fait satisfaisants restent néanmoins à



FIG. 8 – Tennis : Première et deuxième ligne : Images extraites de segments d'échange et de service. Troisième et quatrième ligne : Images extraites d'autres segments (vidéo fournie par l'INA).

interpréter avec prudence. En effet, nous avons beaucoup plus de segments d'échange que de service. Du fait d'une tendance de l'algorithme de segmentation automatique à une certaine sur-segmentation, un bon nombre de segments de service obtenus étant très courts (inférieurs à une seconde), ils ont été systématiquement supprimés et, par conséquent n'apparaissent pas dans cette étude. D'autre part, les segments intrus ne figurent pas non plus dans ce tableau. Nous complétons actuellement la méthode de reconnaissance par l'ajout d'une classe "autre" pour ne pas forcer les intrus issus du tri à être placés dans une des classes d'événements pertinents. Cela passe par l'apprentissage d'une classe "intrus". Les résultats obtenus pour le

	Etiquette attribuée		
	Echange	ServBust	ServEnt
Echange	50	2	0
ServBust	0	5	1
ServEnt	0	0	1

TAB. 4 – Tennis : Résultats de l'étape de reconnaissance (en nombre de segments). Première colonne = vérité terrain.

moment sur cette vidéo sont encourageants, nous allons cependant étendre cette validation sur un corpus plus important de vidéos.

7 Conclusion

Nous avons présenté une approche originale et efficace pour la détection d'événements dans une vidéo, s'appuyant sur l'apprentissage de modèles probabilistes de mouvement. Elle tient explicitement compte des informations liées au mouvement de la scène et au mouvement de la caméra. Nous avons ainsi introduit un modèle probabiliste adapté des mesures locales de mouvement résiduel et de leurs contrastes temporels. Une modélisation probabiliste appropriée du mouvement de la caméra a également été conçue. Le principe de la méthode de détection d'événements pertinents proposée est général, puisqu'il ne requiert pas de connaissance spécifique sur le genre de la vidéo, comme le type de sport considéré par exemple. Le "monde des événements" est bien sûr lui à spécifier pour chaque application puisqu'il forme la "sémantique" apportée. Notre méthode peut par conséquent s'appliquer à un large champ de vidéos où le mouvement représente un aspect déterminant. De plus, bien qu'elle soit supervisée, son déroulement en deux étapes allège considérablement la phase d'apprentissage. Le cadre statistique considéré offre, quant à lui, une flexibilité permettant d'introduire aisément des a priori sur les classes, ou d'ajouter d'autres types d'informations telles que la couleur ou des descripteurs audios extraits de la bande son.

Annexe

Soit $f_{v_{res}}^1(x) = \beta_1 \delta_0(x) + (1 - \beta_1) \phi_t(x; 0, \sigma_1^2) \mathbf{1}_{x>0}$ et $f_{v_{res}}^2(x) = \beta_2 \delta_0(x) + (1 - \beta_2) \phi_t(x; 0, \sigma_2^2) \mathbf{1}_{x>0}$ les densités de distribution des mesures locales de mouvement associées respectivement aux segments s_1 et s_2 . La distance de Kullback-Leibler entre ces deux lois est donnée par :

$$d_K(f_{v_{res}}^1, f_{v_{res}}^2) = \beta_1 \text{Log} \left(\frac{\beta_1}{\beta_2} \right) + (1 - \beta_1) \text{Log} \left(\frac{\sigma_2(1-\beta_1)}{\sigma_1(1-\beta_2)} \right) + \frac{1-\beta_1}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 \right)$$

La distance de Kullback entre les densités $f_{\Delta v_{res}}^1$ et $f_{\Delta v_{res}}^2$ des contrastes temporels pour les deux segments s_1 et s_2 s'écrit de la même façon. D'après la définition de la loi jointe du mouvement résiduel (4), nous avons :

$$d_K(\mathcal{M}_{res}^1, \mathcal{M}_{res}^2) = d_K(f_{v_{res}}^1, f_{v_{res}}^2) + d_K(f_{\Delta v_{res}}^1, f_{\Delta v_{res}}^2)$$

Finalement, la version symétrisée de la distance de Kullback-Leibler, notée d , s'obtient comme suit :

$$d(\mathcal{M}_{res}^1, \mathcal{M}_{res}^2) = \frac{1}{2} (d_K(\mathcal{M}_{res}^1, \mathcal{M}_{res}^2) + d_K(\mathcal{M}_{res}^2, \mathcal{M}_{res}^1))$$

Remerciements

Ce travail a été en partie financé par la Région Bretagne et par le Ministère de l'Industrie dans le cadre du projet RNTL Domus Videum. Nous remercions également l'INA qui nous a fourni le corpus vidéo utilisé pour les expérimentations.

Références

- [1] M. Basseville. Detecting changes in signals and systems- a survey. *Automatica*, 24:309–326, 1988.
- [2] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(3):719–725, 2000.
- [3] P. Bouthemy, M. Gelgon, and F. Ganansia. An unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 9:1030–1044, 1999.
- [4] E. Bruno, and D. Pellerin. Video structuring, indexing and retrieval based on global motion wavelet coefficients. *IAPR Int. Conf. on Pattern Recognition, ICPR'2002, Québec, Canada*, 3:30287–30290, Août 2002.
- [5] R. Fablet and P. Bouthemy. Non parametric motion recognition using temporal multiscale Gibbs models. *IEEE Int. Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii*, Décembre 2001.
- [6] R. Fablet, P. Bouthemy, and 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.
- [7] J.P. Lecoutre and P. Tassi. *Statistique non paramétrique et robustesse*. Economica, 1987.
- [8] Y-F. Ma and H-J. Zhang. Motion pattern-based video classification retrieval. *EURASIP Journal on Applied Signal Processing*, 2:199–208, Mars 2003.
- [9] J. Nam and H. Tewfik. Dynamic video summarization and visualization. *7th ACM International Conference on Multimedia, ACM Multimedia'99, Orlando*, pages 53–56, Novembre 1999.
- [10] C-W. Ngo, T-C. Pong, and H-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Trans. on Multimedia*, 4(4):446–458, Décembre 2002.
- [11] J-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Communication and Image Representation*, 6(4):348–365, Décembre 1995.
- [12] N. Peyrard and P. Bouthemy. Detection of meaningful events in videos based on a supervised classification approach. *IEEE Int. Conf. on Image Processing, Barcelone*, Septembre 2003.
- [13] Y. Rui and 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.
- [14] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, Janvier 2000.
- [15] Y. Yacoob and J. Black. Parametrized modeling and recognition of activities. *Sixth IEEE Int. Conf. on Computer Vision, Bombay, India*, pages 120–127, 1998.
- [16] L. Zelnik-Manor and M. Irani. Event-based video analysis. *IEEE Int. Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2:123–130, Décembre 2001.