Plateforme pour la détection générique d'évènements suspects dans les enregistrements de vidéosurveillance

Bouchra ABBOUD,

FACING-IT SAS, 7 bis rue du Bataillon de France, 60200

bouchra.abboud@facing-it.net
www.facing-it.net

Résumé – La généralisation des caméras de surveillance et la multiplication des données vidéo disponibles a généré le besoin pour des outils d'indexation vidéo génériques rapides et fiables capables d'effectuer une recherche semi-automatique d'évènements dans une base de données vidéo. Dans cet article nous proposons une plateforme générique pour la détection d'évènements dans les enregistrements de vidéosurveillance basée sur un apprentissage statistique des évènements spatio-temporels couplé à un module de active learning permettant de tenir compte du feedback de l'utilisateur. Cette technique de représentation non paramétrique présente l'avantage indéniable de gérer des évènements divers et a été testée avec succès pour la segmentation temporelle et le clustering de séquences vidéo

Abstract – Large quantities of video surveillance data exist today due to the wide spread of cameras. The main use of this video data is to review tape after a known event and gather information about a specific event. This task could be made easier for the human analyst if the video recordings could be previously indexed and segmented as normal or suspicious using a semi-automatic computer program. This paper describes a generic semi-automatic platform event detection in a video recording based on weakly supervised statistical modelling and active learning.

1. Introduction

1.1 Contexte

La généralisation des caméras de vidéosurveillance et la multiplication des données numériques disponibles a crée le besoin pour des outils d'indexation rapides et fiables, permettant aux utilisateurs finaux d'accéder facilement aux informations recherchées dans une base de données vidéo. Cette problématique suscite à ce jour l'intérêt de nombreux chercheurs et fait l'objet de plusieurs projets de recherche (MUSCLE, CARETAKER) et campagnes d'évaluation (ImageEval, TRECVID).

Pour présenter un intérêt auprès d'utilisateurs finaux, tout système d'indexation vidéo doit respecter un certain nombre de contraintes dont la rapidité, la généricité et la facilité d'utilisation. Nous proposons dans cet article une technique de détection d'évènements générique basée sur un apprentissage statistique couplé à une boucle de relevance feedback. L'objectif étant de permettre d'entraîner le détecteur sur divers types d'évènements en présentant un nombre significatif d'exemples dans une approche proche de celle décrite en [1], et ce pour un souci de généricité. Par la suite, le relevance feedback permet d'affiner la sortie du détecteur pour le rendre plus précis de manière intuitive et sans solliciter d'expertise particulière de la part de l'utilisateur final.

Des exemples d'application de cette technique sont la vidéosurveillance intelligente des aéroports et centres commerciaux pour la détection automatique d'évènements tels que la disparition d'objets ou la détection d'évènements susceptibles de constituer une menace en vue de les signaler à un opérateur humain (personne abandonnant un colis par exemple).

1.2 Travaux antérieurs

L'indexation multimédia constitue un défi dans le sens où il s'agit intrinsèquement de combler le fossé sémantique existant entre les propriétés visuelles (texture, couleur, etc.) et le contenu sémantique des vidéos.

Dans ce cadre, une première approche implicite consiste à extraire à partir des données vidéo des attributs spatio-temporels suffisamment compacts et fiables, permettant de représenter de manière efficace les évènements qui se produisent dans la séquence vidéo. Pour ce faire, de nombreuses approches sont proposées dans la littérature dont nous citerons notamment l'extraction d'attributs colorimétriques spatio-temporels [2,3] combinée à des techniques de modélisation statistique telles que les modèles de Markov cachés [4,5]. Une modélisation par noyaux est également utilisée dans [6] pour construire une modélisation d'objets recherchés.

Une seconde approche consiste à lier de manière explicite les attributs visuels au contenu sémantique des séquences à l'aide notamment d'une détection et d'un suivi des objets en mouvement [7,8] couplés à une reconstruction des trajectoires [9].

Cet article décrit une approche basée sur une projection temporelle des évènements [10] couplée à une mise en correspondance pour une détection générique d'évènements par active learning.

FIG. 1 : Exemple d'évènement suspect dans la base de CAVIAR : Une personne abandonne un colis [11]

2. Représentation d'un évènement

La représentation idéale d'un évènement consiste à calculer une projection identique pour un même évènement capturé par des capteurs différents avec des conditions d'éclairage, d'orientation et à des échelles différentes.

Dans cet article nous proposons de considérer un évènement par un tenseur en trois dimensions (x,y,t) également connu sous le nom d'objet temporel. Cette représentation présente des avantages indéniables pour la détection d'évènements par des capteurs différents. En effet, la projection des évènements sur l'espace temporel est invariante à l'échelle de l'image et à la position du capteur. En l'occurrence, le nombre de trames requis pour représenter un évènement est fixe quelle que soit la position, l'orientation ou la résolution du capteur utilisé pour le filmer.

Pour calculer les attributs d'un évènement un sous échantillonnage temporel de la séquence correspondante est effectué à 4 niveaux différents (1 trame sur 2, une trame sur 3, une trame sur 4, et une trame sur 5). Pour chaque niveau de la pyramide temporelle le gradient local d'intensité est calculé pour chaque direction (x,y,t).

Ce gradient ($\mathbf{G}_x^l$, $\mathbf{G}_y^l$, $\mathbf{G}_t^l$) représente la normale à la surface spatio-temporelle générée par la représentation de l'évènement à l'échelle l de la pyramide temporelle.

D'une part, la direction du gradient constitue une représentation de l'orientation de la surface spatio-temporelle et donc du mouvement global des objets dans l'évènement. D'autre part, la norme du gradient dépend essentiellement de l'apparence (texture et couleur) des trames constituant l'évènement.

Pour calculer une représentation de l'évènement la plus indépendante possible de l'apparence et de la couleur des images, la norme du gradient est normalisée à une valeur unité. De même la direction du gradient est uniquement considérée dans sa valeur absolue pour ne pas tenir compte de la direction du mouvement (personne se déplaçant à gauche versus à droite par exemple).

$$\left(\overline{G_x^l, G_y^l, G_t^l}\right) = \frac{\left(|G_x^l|, |G_y^l|, |G_t^l|\right)}{\sqrt{(G_x^l)^2 + (G_y^l)^2 + (G_t^l)^2}} \qquad (1)$$

Pour chaque point spatio-temporel (x,y,t) la distribution statistique de la direction du gradient normalisé est calculée pour chaque échelle « l » de la pyramide temporelle.

Un évènement est donc représenté par 12 histogrammes qui correspondent aux gradients spatiaux et temporels normalisés du tenseur vidéo sous échantillonné à 4 échelles temporelles différentes. En l'occurrence, une échelle temporelle correspond à un sous échantillonnage des trames de la vidéo.

Ces histogrammes correspondent à la distribution des changements spatio-temporels générés par l'évènement au cours de la scène. Contrairement aux attributs basés sur l'apparence, ces attributs spatio-temporels sont peu sensibles aux changements de couleur, d'arrière plan, d'angle de prise de vue et de taille.

De même, les histogrammes ainsi calculés sont lissés et leurs intégrales respectives sont normalisées à l'unité. L'objectif de cette normalisation est de compenser les variations dues à des évènements similaires dont les échelles spatiales ou temporelles sont différentes. Ceci concerne en l'occurrence les évènements similaires filmés à des résolutions différentes et/ou dont les durées ne sont pas égales.

Pour vérifier la capacité des attributs ainsi calculés à représenter des évènements semblables captés dans des directions différentes avec des attributs colorimétriques différents les tests suivants sont réalisés [10].

D'une part, l'évènement « walks » réalisé par des personnes différentes et habillées avec des vêtements différents est filmé dans trois orientations différentes avec trois capteurs différents et à des angles de vue différents (figure 2). Les attributs spatio-temporels de ces trois évènements sont reportés sur la figure 4 en bleu. D'autre part, les attributs temporels des évènements, « jog », « wave » et « roll » sont également calculés sur une seule séquence chacun (figure 3) et sont reportés sur la figure (4) dans des couleurs différentes. La figure (3) permet de constater visuellement que les attributs spatio-temporels des trois évènements « walks » sont visuellement regroupés et facilement séparables des attributs relatifs aux autres

évènements. Ceci permet de penser que la représentation par attributs spatio-temporels proposée permet effectivement de discriminer avec succès les évènements et constitue un sous-espace robuste aux variations de couleur, d'orientation d'échelle et de durée.

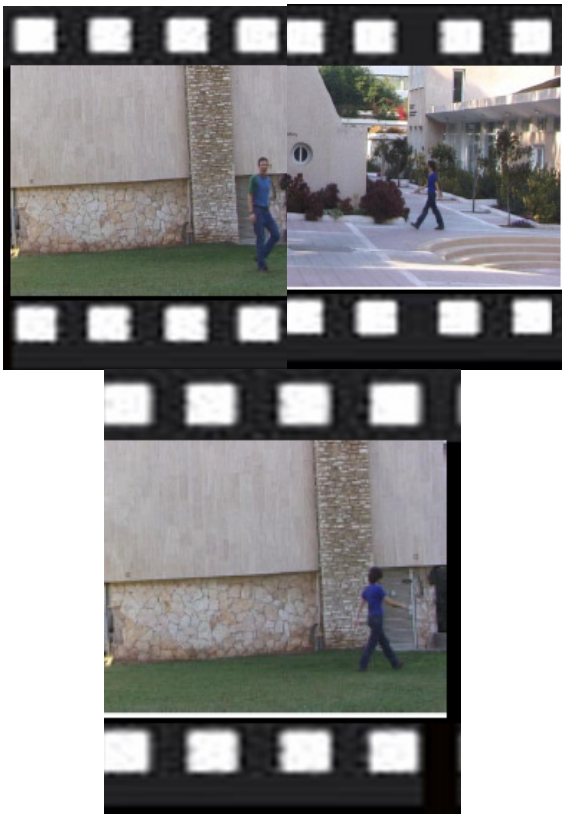

FIG. 2 : Capture de l'évènement « walks » dans des directions différentes avec des arrières plans différents [10]

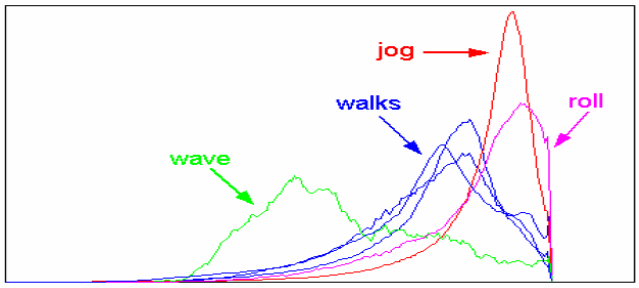

FIG. 4: Les attributs spatio-temporels de l'évènement « walks » capturé avec trois arrière plans différents et dabs des directions différentes sont visuellement regroupés et séparables des attributs relatifs à d'autres évènements [10]

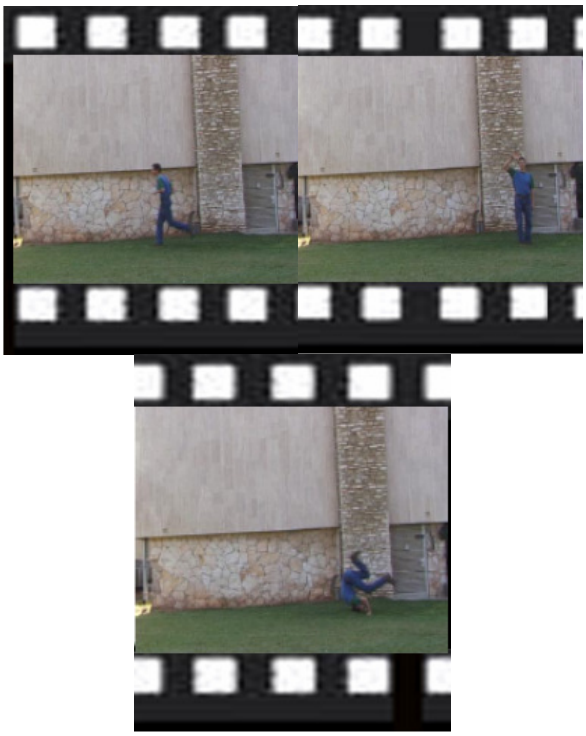

FIG. 3 : Capture des évènements « jog », « wave » et « roll » [10]

3. Détection d'un évènement

Ayant défini dans la section précédente un espace d'attributs spatio-temporels, permettant de représenter les évènements de manière compacte, fiable et robuste aux variations de couleur et d'orientation, il convient à présent de définir les métriques permettant de mettre en correspondance deux représentations de manière à reconnaître un évènement particulier. Dans le cas précis de cet article nous proposons une plateforme générique basée sur un apprentissage statistique de l'évènement recherché par la présentation de N échantillons vidéo représentant cet évènement. La disponibilité de ces N échantillons permet donc d'envisager de mutualiser leurs apports dans la représentation spatio-temporelle de l'évènement.

Pour ce faire, une première approche consiste à calculer la moyenne des attributs extraits à partir de chacun des échantillons puis à utiliser cette moyenne pour représenter l'évènement. Cette approche peut s'avérer intéressante ou nécessaire, lorsque le nombre d'échantillons disponibles est très faible.

Cependant, lorsqu'un nombre suffisant d'échantillons est disponible l'utilisation d'une métrique plus évoluée tenant compte de la variabilité des échantillons peut s'avérer très intéressante. Nous retiendrons cette dernière approche.

Pour détecter un évènement particulier dans un enregistrement vidéo une première étape consiste donc à présenter N exemples de séquences représentatives de l'évènement et à calculer les attributs spatio-temporels

$\mathbf{h}_{k=1...,N}$ des évènements capturés sur chacune de ces séquences.

Une deuxième étape consiste à mettre en correspondance les attributs ainsi obtenus avec ceux extraits à partir de la séquence analysée et à dessiner la courbe de probabilité d'occurrence de l'évènement appris durant le déroulement de la vidéo.

Pour ce faire une fenêtre glissante de longueur T est appliquée à la vidéo traitée et les attributs spatio-temporels $\mathbf{h}_i$ correspondants sont calculés pour chaque sous séquence « i ».

La similitude entre l'évènement recherché représenté par les N vecteurs d'attributs spatiaux temporels $\mathbf{h}_{k=1,...,N}$ et la fenêtre glissante « i » est estimée à l'aide de la distance de Mahalanobis

$$d^2(i,k) = (h_i - \overline{h_k})\Sigma_{h_k}^{-1}(h_i - \overline{h_k})^{-1} \qquad (2)$$

Les fenêtres glissantes qui possèdent la distance la plus proche à l'évènement recherché sont alors présentées à l'utilisateur final et ce dernier est invité à sélectionner parmi ces dernières celles qui lui semblent les plus pertinentes. La matrice de covariance et la moyenne des attributs spatio-temporels sont alors recalculées à l'aide de ce nouvel ensemble d'apprentissage manuellement enrichi par l'utilisateur et les fenêtres les plus vraisemblables sont recalculées et présentées par ordre de ressemblance au nouvel ensemble d'apprentissage. Ce processus est répété jusqu'à ce que les résultats soient jugés satisfaisants par l'utilisateur ou jusqu'à la stabilité du retour [12].

4. Conclusion

Nous avons présenté une plateforme générique pour la détection d'évènements dans des enregistrements vidéo. Cette technique possède des applications très vastes pour l'indexation des enregistrements de vidéo surveillance et la détection d'évènements précis (intrusion humaine dans une zone interdite, personne prenant ou abandonnant un colis, etc.) dans une base de données vidéo. L'idée principale de cette plateforme consiste à apprendre l'évènement désiré à l'aide d'un nombre d'exemples pertinents présentés par l'utilisateur et à proposer dans un premier temps les séquences les plus « ressemblantes ». L'utilisateur est ensuite invité à sélectionner manuellement les retours qui lui semblent les plus pertinents et la recherche est relancée avec ce nouvel ensemble d'apprentissage. Cette technique également connue sous le nom de active learning permet de construire des détecteurs génériques (pouvant détecter des évènements divers) et de les entraîner avec peu d'exemples au départ.

Les résultats préliminaires obtenus grâce à cette méthode sont très encourageants, des améliorations notables des performances du système sont également attendues notamment par l'utilisation de méthodes d'apprentissage et de mise en correspondance plus évoluées tels que les SVM[13] ou les HMM[14]. Cependant, ceci nécessite de présenter des exemples « négatifs » en même temps que les exemples « positifs » lors de la phase d'apprentissage rendant le système moins simple d'utilisation.

Références

[1] Y. Ke,, R. Sukthankar and M. Herbert, *Event Detection in Crowded Videos*, International Conference on Computer Vision (ICCV07), Rio de Janeiro, Brazil, 2007

[2] G. Lavee, L. Khan and B Thuraisingham, *A Framework for a Video Analysis Tool for Suscpicious Event Detection*, Multimedia Tools Appl. 35(1), 109-123, 2007

[3] M. Chen, S. Chen and M. Shyu, Hierarchical Temporal Association Mining for Video Event Detection in Video Databases, International Conference on Data Engineering Workshop, Istanbul, 2007

[4] A. Bobick, A. Pentland and T. Poggio, *Learning and Understanding Action in Video Imagery*, Proceedings DARPA Image Understanding Workshop, 1998.

[5] G. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. Strintzis, *Estimation and Representation of Accumulated Motion Characteristics for Semantic Event Detection*, International Conference on Image Processing (ICIP), San Diego, California, USA, 2008.

[6] P.H. Gosselin, M. Cord, and S. Philipp-Foliguet, *Kernel on Bags for Multi-Object Database Retrieval*, ACM International Conference on Image Processing, Atlanta USA, 2006

[7] G. Medioni, I. Cohen, F. Brémond, S. Hongeng and R. Nevatia, *Event Detection and Analysis from Video Streams,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol23, 2001.

[8] I. Laptev, B. Capuo, Ch. Schultz and T. Lindeberg, *Local Velocity-Adapted Motion Events for Spatio-Temporal Recognition*, Computer Vision and Image Understanding, 108(3):207-229, 2007.

[9] Projet ANR KIVAOU, AAP CSOSG 2007.

[10] L. Zelnik-Manor and M. Irani, *Event Based Analysis of Video*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[11] Projet CAVIAR Information Society Technology's programme project IST 2001 37540.

[12]M. Cord and P.H. Gosselin, *Online Context-Based Image Retrieval Using Active Learning*, In Machine Learning Techniques for Multimedia, chapter 5, Springer, 2008.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995

[14] L. R. Rabiner *A tutorial on Hidden Markov Models and selected applications in speech recognition,* Proceedings of the IEEE 77 (2): 257-286, 1989.