Intégration de la saillance visuelle dans la reconnaissance d'évènements rares

Dieudonné Fabrice ATREVI[1], Damien VIVET[2], Bruno EMILE[1]

[1]Université d'Orléans, INSA Centre Val de Loire, PRISME EA 4229, F45072, Orléans, France

[2]ISAE-Supaero, Université de Toulouse, Toulouse, France

fabrice.atrevi@univ-orleans.fr, damien.vivet@isae.fr
bruno.emile@univ-orleans.fr

Résumé – Cet article présente une nouvelle méthode de détection d?événements rares pouvant être qualifier d'anormaux dans une vidéo. Elle s'appuie d'une part sur l'utilisation de la saillance visuelle et d'autre part sur la détection et la description locale des points d'intérêts. Un filtrage des points d'intérêt est effectué grâce au score de saillance permettant ainsi de ne prendre en considération que ceux ayant un impact visuel important. Un modèle d?événements normaux est appris grâce à l'utilisation d'un modèle génératif probabiliste "allocation de Dirichlet latente" (LDA), connu pour sa performance en fouille de données textuelles. La détection d'un événement anormal ou rare est alors effectuée de façon probabiliste via le modèle appris. Nous proposons dans cet article, de combiner une focalisation visuelle par utilisation de la saillance et l'utilisation d'une technique de classification automatique de documents afin de classifier les images d'une vidéo et ainsi détecter les événements rares.

Abstract – This paper presents a new method for the detection of rares events in video. It is based on the visual saliency and on the detection and local description of points of interest. The point-of-interest filtering is carried out using the saliency score, allowing only those with visual importance to be considered. A model of normal events is learned thanks to the probabilistic generative model "Latent Dirichlet Allocation" (LDA), known for its performance in textual data mining. The detection of an abnormal or rare event is carried out in a probabilistic way via the learned model. This paper proposes to combine a saliency based visual focalization and the use of automatic document classification technic in order to classify images from a video and to detect rare events.

1 Introduction

Afin d'accroître la sécurité des lieux publics, la question de la mise en place de système automatique de détection d'événements anormaux a été abordée par la communauté de vision par ordinateur depuis des dizaines d'années. Le but étant de mettre au point des algorithmes fiables pouvant détecter des faits anormaux ou rares dans une vidéo. Par événements rares, nous entendons, tout comportement déviant d'un ensemble d'observations considérées comme normales. En fonction de l'application visée, cette notion d'anormalité varie, ce qui constitue un des défis majeurs. De nombreuses approches ont ainsi été proposées [1] [2][3]. La plupart d'entre elles se basent sur le calcul, soit du flot optique, soit du gradient spatio-temporel, soit sur une combinaison de l'histogramme des gradients orientés et celui de flot optique orienté. Ces différents descripteurs sont généralement calculés dans l'entourage de points d'intérêt extrait des images. D'autres travaux proposent la modélisation des événements en ayant recours à des modèles de chaînes cachées de Markov, de mixture de gaussiennes ou de modèles basé sac de mots.

Les approches se basant sur la détection de points d'intérêt et leur description, nécessitent non seulement une bonne détection de ces points, mais également un choix pertinent de ceux qui décrivent le mieux les objets présents dans la scène. Turcot et Lowe [4] ont affirmé que le choix pertinent d'un petit nombre de caractéristiques est mieux qu'un grand nombre pour des problèmes de reconnaissance. C'est fort de ce résultat que nous pensons que l'utilisation de méthodes de saillance visuelle pour la sélection intelligente des points d'intérêt permettra une amélioration de la description des mouvements dans une scène. Un tel choix peut également être justifié afin d'éviter la détection de points d'intérêt sur des objets ne générant pas d ?événements notables dans la scène, à savoir les points appartenant au fond. Des méthodes de sélection de points pertinents ont été proposées dans la littérature. Laptev et al. [5] propose une extension des points Harris 2D en prenant en compte l'aspect temporel par un filtrage gaussien temporel des points Harris 2D. Chakraborty et al. [6] propose également une méthode de sélection de points Harris 2D constituée d'une série d'opérations visant à attribuer à chaque point un facteur de poids lié à son voisinage. Une phase de suppression prenant en compte ce poids et des contraintes spatio-temporelles permet de filtrer les points Harris 2D. Aucune de ces méthodes, bien que présentant des résultats acceptables, n'intègre l'information de saillance visuelle comme critère de filtrage. Nous avons ainsi choisi d'investiguer cette forme de filtrage pour une intégration

dans un modèle de détection d'événements rares. Ce modèle de détection est basée sur une approche de classification non supervisée.

2 Approche proposée

La méthode que nous proposons dans cet article est une approche non-supervisée permettant de reconnaître des événements rares, donc ne se produisant pas régulièrement, dans une vidéo. La figure 1 présente le schéma général de notre approche. Nous présenterons dans les sous-sections suivantes, chacune des composantes de la méthode.

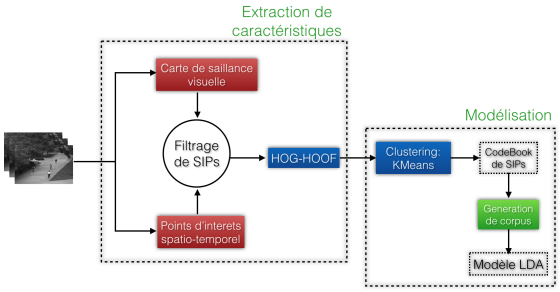

FIGURE 1 – Schéma de la méthode

2.1 Calcul de la saillance visuelle

Un objet est dit saillant s'il peut être facilement remarqué par la vision humaine. En partant de cette idée, les algorithmes développés en saillance visuelle ont pour objectif de mettre en évidence différents objets susceptibles d'attirer l'attention humaine dans une image.

L'objectif de notre approche est de partir d'une répartition des probabilités de saillance de chaque point d'intérêt pour faire un filtrage et ne retenir que ceux ayant un score suffisant pour être considérés comme pertinents dans la scène. L'idée derrière ce choix est que les événements sont générés en grande partie par des objets saillants en mouvement (piétons, voiture, etc.) ou tout du moins n'appartenant pas au fond. On peut ainsi, en extraillant les objets saillant dans le temps, arriver à mieux caractériser l'événement en cours dans la vidéo. L'algorithme de saillance utilisé dans le cadre de nos travaux est celui proposé par Xiaodi Hou et al [7]. Celui-ci se base sur le calcul de la signature (confère formule 1) de l'image dans la base des cosinus discrets ("Discret Cosinus Transform").

$$ImageSignature(x) = sign(DCT(x)) \quad (1)$$

La carte de saillance est obtenue par l'équation 2 ci-après :

$$m = g * (\bar{x} \circ \bar{x}) \quad (2)$$

où $\bar{x}$ représente l'image reconstruite à partir de sa signature, g représente une fonction gaussienne et $\circ$ l'opérateur Hadamard. La figure 2 présente un résultat de l'application de l'algorithme sur une image de la base UCSD Ped1 [1].

On peut remarquer que la majorité des personnes en mouvement (soit les zones saillantes d'intérêt) sont captées par l'algorithme.

1. http://hci.iwr.uni-heidelberg.de/COMPVIS/research/abnormality

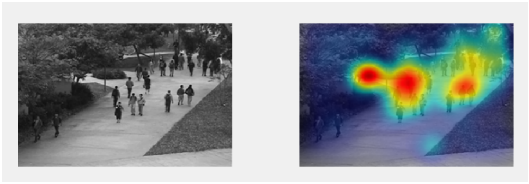

FIGURE 2 – Carte de saillance, les piétons en mouvement sont bien captés par l'approche utilisée.

2.2 Extraction et Description des points d'intérêts

Dans cette étape, nous avons procédé à l'extraction de points d'intérêt Harris 2D. Ces détections sont alors réparties sur l'ensemble des personnes, voitures et différents objets présents dans la scène. Tous ces objets n'étant pas important à la caractérisation des événements se produisant dans une vidéo, nous proposons alors de les sélectionner en nous basant sur leur probabilité de saillance. La sélection consiste alors à fixer un seuil de probabilité. Tous les points ayant une probabilité inférieure à ce seuil sont éliminés. On peut observer un résultat de seuillage sur la figure 3.

FIGURE 3 – Résultat de sélection de points d'intérêt par utilisation de la probabilité de saillance. A gauche la détection initiale et à droite détection après filtrage

La description des points ainsi sélectionnés est réalisée à partir de l'histogramme des orientations du gradient pour prendre en compte l'information d'apparence et celui des orientations du flot optique pour l'information de mouvement. Ils sont extraits à partir de la région centrée sur les points d'intérêt obtenus après filtrage. Les deux histogrammes sont par la suite concaténés pour obtenir le descripteur final devant caractériser chaque point.

2.3 Modélisation des événements normaux

Une fois nos descripteurs extraits, nous procédons à leur modélisation par le biais d'une méthodologie inspirée de l'approche "sac de mots" et du modèle non-supervisé "Latent dirichlet allocation" [8]. Cette méthode présente l'avantage de ramener dans un même espace, toutes images ou parties d'images, quels que soient le nombre de points d'intérêt qu'elles contiennent. Les vecteurs de descripteurs issus de cette étape sont généralement entraînés par des algorithmes d'apprentissage tels que les machines à supports de vecteurs (SVM). Dans notre approche, nous nous proposons de remplacer l'utilisation des algorithmes tels que les SVM par un algorithme non-supervisé, qui offre l'avantage de ne pas nécessiter des données labellisées (dans

notre cas des exemples d'événements normaux et anormaux). La liste des événements anormaux étant exhaustive, il est difficile de constituer une base de ces derniers, d'où la nécessité d'utiliser des algorithmes non-supervisés.

2.3.1 Mise en place des sacs de mots

Dans cette étape, qui est l'une des plus cruciales, nous avons utilisé l'algorithme de clustering KMeans, pour dégager un certain nombre de groupe de points d'intérêts. Leurs centroïdes seront considérés comme les mots de notre vocabulaire. Ces différents groupes peuvent représenter des points décrivant des objets aux propriétés similaires (apparence, mouvement).

Une fois les mots du dictionnaire obtenus, il faut pour chaque point, déterminer le mot dont il est le plus proche. Cela se fait en calculant la distance euclidienne, qui mesure la similarité, entre le point et tous les mots du vocabulaire. Ainsi, chaque point votera pour un mot, ce qui nous permettra d'avoir un vecteur d'occurrence des mots qui représentera l'image. Il est à noter que le nombre de mots dépend de la richesse des scènes et des objets à décrire.

2.3.2 Latent Dirichlet Allocation

Le modèle LDA (Latent Dirichlet Allocation), connu pour ses performances en fouille de texte et en classification d'objets, est un modèle probabiliste génératif qui permet de décrire des collections de documents de texte ou d'autres types de données discrètes. Cette description consiste à dégager des thèmes traités par un corpus en se basant sur les mots qui le compose. Les terminologies utilisées pour l'interprétation ce modèle sont plus en rapport avec les textes. Nous présentons ci-dessous leurs équivalences pour nos jeux de données.

- Un mot w est la plus petite unité de données, correspondant à l'indice d'un mot dans un vocabulaire fixe de taille V. Dans notre cas, les mots du vocabulaire sont les centroïdes issus de l'algorithme de clustering et donc un mot correspond au vecteur de descripteur (dans notre cas HOG-HOOF) associé à un point d'intérêt.

- Un document est constitué d'un ensemble de mots. Il s'agit dans notre cas d'une image. Plus concrètement, il s'agit d'un vecteur contenant le nombre d'occurrences de chaque mot du vocabulaire dans l'image.

- Un corpus est une collection de documents. Celui d'apprentissage représente l'ensemble des vecteurs d'occurrences de mots dans toutes les images de la base d'apprentissage.

- Les topics sont découverts lors du processus d'apprentissage. Ils permettent de mettre en évidence, les mots qui apparaissent souvent ensemble pour décrire un même fait. Dans notre méthode, nous imaginons que les topics permettront de mettre en perspective des événements. Une distribution de ces topics pourra permettre de caractériser les événements normaux et anormaux. Des détails sur le processus de génération des topics et d'optimisation des paramètres peuvent être consulté dans [8].

3 Experimentations

Dans cette section, nous présentons les expérimentations qui ont été effectuées sur la base UCSD Ped 1. Elle est composée d'un ensemble de 34 vidéos d'apprentissage, contenant uniquement des images d'événements normaux et d'un autre ensemble de 36 vidéos de test contenant quelques images d'événements anormaux. De façon globale, les images d'événements normaux représentent des scènes d'une zone piétonnes où ne passent uniquement que des piétons. Les événements anormaux sont des séquences contenant des voitures, motos qui passent également par cette zone piétonne.

Les paramètres de notre modèle ont été déterminés après un certain nombre de tests avec différentes valeurs. Nous avons fixé, pour les résultats présentés ci-dessous, le seuil de filtrage des points d'intérêt à 0.3, le nombre de mots de notre vocabulaire à 1000, le nombre de topics à découvrir par le LDA à 30, la taille des vecteurs de descripteur à 72 pour HOG et 32 pour HOOF.

Deux types d'expériences ont été menées sur les différentes vidéos. Le premier type d'expérience a été réalisé afin d'observer le fonctionnement du modèle face à ce jeu de données complexe et donc pour comprendre les topics identifiés par l'algorithme LDA. Le second type d'expérience consiste à analyser le comportement du modèle dans le cadre d'une détection d'événements anormaux.

3.1 Interprétation de topics

En fouille de données textuelle, il est plus ou moins facile d'interpréter les topics identifiés par le LDA. Par contre, dans le cas d'analyse de scènes, cette tâche est plus complexe. Il est néanmoins nécessaire de comprendre les topics identifiés par l'algorithme afin d'être sûr de la pertinence de l'information extraite. Des séries de tests ont été menées dans le but d'identifier, pour chaque image d'une vidéo, le topic qui est le plus représenté. Il est alors possible d'avoir une statistique des topics présents dans chaque vidéo et ainsi de donner une interprétation de ces derniers.

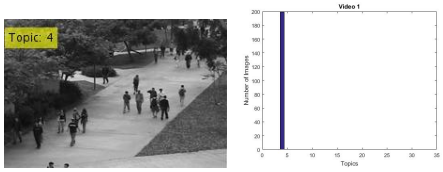

FIGURE 4 – Classification de scène : L'ensemble des images de la vidéo 1 appartiennent au topic 4.

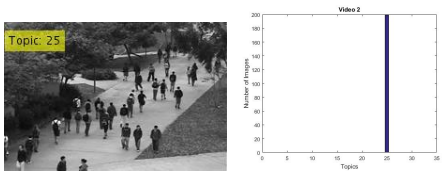

FIGURE 5 – Classification de scène : L'ensemble des images de la vidéo 2 appartiennent au topic 25.

Les figures 4 et 5 présentent deux différents topics identi-

fiés dans deux différentes vidéos. On peut noter que toutes les images de ces vidéos appartiennent au même topics respectivement 4 et 25. Nous pouvons observer que la méthode permet de mettre en évidence les images présentant la même densité de piétons. Dans certaines vidéos dans lesquelles cette densité varie, on constate également une variation des topics. Cependant, nous ne pouvons pas conclure que les topics ainsi identifiés pourront servir à faire la différence entre des images d'événements normaux et anormaux. Ici nous pouvons néanmoins assez nettement détecter le niveau d'affluence dans la zone observée (une forte affluence pourrait être considérée comme anormale selon l'application visée).

3.2 Classification d'évènements

Pour analyser les performances de notre méthode, il est important de tracer la courbe ROC du modèle de classification. Nous avons donc procédé à des expérimentations sur la base de test dont la vérité terrain est disponible. La sortie du modèle est la log-vraissemblance pour chaque image au regard du modèle d'événements normaux appris. Celle-ci est obtenue par la formule 3 avec α et β des hyperparamètres du modèle et D une image.

$$l(\alpha, \beta) = \log p(D|\alpha, \beta) \tag{3}$$

La tracé de la courbe ROC est réalisée en variant le seuil sur le score de log-vraissemblance.

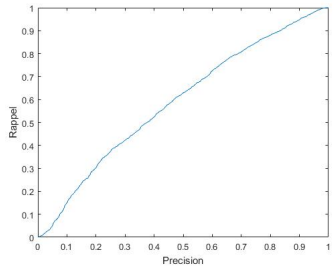

FIGURE 6 – Courbe ROC sur la base de Test UCSDPed 1

La courbe de la figure 6 représente le résultat obtenu de notre modèle et renseigne que ce dernier éprouve des difficultés à séparer les images d'événements normaux de ceux anormaux. La surface sous la courbe (AUC) obtenue est de 0.5980 qui est sensiblement égale à celui de Mixture Dynamic Textures (MDT) spatial [2], mais reste encore loin des meilleures performances obtenues sur cette même base. Cette différence réside dans le fait que notre méthode considère l'image dans sa totalité. Ainsi, tous les points d'intérêt d'une image participent à la représentation générale de l'image et non pas uniquement à une fenêtre de l'image. Les événements anormaux étant très localisés dans la scène (apparition d'un vélo dans la foule), notre approche en l'état n'est pas capable de se focaliser sur cet événement local. De plus, l'approche proposée s'appuie sur une classification image par image sans prendre pour l'instant en compte l'aspect temporel. Ces deux modifications (prise en compte de l'aspect temporel, et spatialisation du traitement) sont envisagées comme perspectives à ce travail.

4 Conclusion

Nous avons présenté dans cet article, une approche intégrant la saillance visuelle pour le choix de points d'intérêt et la modélisation de scène avec l'algorithme d'Allocation Dirichlet Latent pour la classification d'événements dans une vidéo. Les premiers résultats obtenus de la méthode proposée sont prometteurs en ce sens qu'elle permet de capter l'information concernant le mouvement des entités présentes dans la scène. Elle permet notamment de détecter aisément la densité d'objets mobiles présents dans la scène. Notre approche permet également de classifier certains événements anormaux avec un score AUC de 0,5980 en prenant l'image dans son ensemble et sans considérer l'aspect temporel. Les perspectives de ce travail consisteront à prendre en compte l'aspect temporel dans le calcul de la saillance et des descripteurs de points d'intérêt tout comme l'analyse locale de zones dans l'image pour détecter et sélectionner les événements très localisés de la scène.

5 Remerciements

Ce travail fait partie du projet LUMINEUX, soutenu par la Région Centre-Val de Loire (France). Les auteurs tiennent à remercier le Conseil Régional du Centre - Val de Loire pour son soutien.

Références

[1] Dinesh Singh and C Krishna Mohan. Graph formulation of video activities for abnormal activity recognition. *Pattern Recognition*, 65 :265–272, 2017.

[2] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[3] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. Learning deep event models for crowd anomaly detection. *Neurocomputing*, 219 :548–556, 2017.

[4] Panu Turcot and David G Lowe. Better matching with fewer features : The selection of useful features in large database recognition problems. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 2109–2116. IEEE, 2009.

[5] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3) :107–123, 2005.

[6] Bhaskar Chakraborty, Michael B Holte, Thomas B Moeslund, and Jordi Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3) :396–410, 2012.

[7] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature : Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, 34(1) :194–201, 2012.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.