Production d'annotations par plan pour l'indexation des vidéos

Nadia Derbas [1]

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

{Prénom.Nom}@imag.fr

RÉSUMÉ. *La qualité des annotations dans les vidéos d'entraînement joue un rôle très important dans la qualité des systèmes de détection automatique d'événements dans les vidéos. Dans cet article, nous proposons une méthode pour la génération d'annotations au niveau des plans à partir d'annotations au niveau des vidéos complètes. Cette méthode utilise des techniques de filtrage en fonction du contenu visuel des vidéos et elle est basée sur l'idée que les plans contenant un événement donné ont tendance à être semblables entre eux contrairement aux plans ne contenant pas cet évènement, qui auront plus tendance à être différents du reste des plans. La méthode a été implémentée et évaluée dans le cadre de la tâche de détection d'événements multimédias (MED) à TRECVID 2011. Les expérimentations montrent que l'approche ne permet pas toujours une amélioration mais produit un gain pour certains événements.*

ABSTRACT. *The quality of the annotations in the training videos has a very important role in the quality of automatic systems for event detection in videos. This paper presents a method for the generation of shot level annotations from whole video level annotations. This method uses a filtering technique based on the visual content and on the idea that shots containing a target event are similar while shots not containing it are different from each other and from the plans which contain this event. The method was implemented and evaluated on the Multimedia Event Detection task (MED) of TRECVID 2011. The experiments show that this approach does not always improve the result of event detection although it provides some gain for some specific concepts.*

MOTS-CLÉS : *Filtrage d'information, contenu visuel, indexations de vidéos*

KEYWORDS: *Information filtering, visual content, video indexing*

1. *Sous la direction de Georges Quénot*

1. Introduction

L'explosion du volume d'images et de vidéos durant ces dernières années a créé un grand besoin pour la recherche d'information dans les documents multimédia à partir du contenu. La diversité des vidéos et leurs environnements multiples rendent la reconnaissance d'objets et l'estimation du mouvement très difficiles pour les systèmes d'indexation de vidéos. Néanmoins, la détection automatique d'événements dans les vidéos d'Internet possède un grand potentiel dans différentes applications comme la recherche de vidéos en général, de gestion de contenu, etc.

Une grande proportion des vidéos d'Internet sont des vidéos « amateurs », faites et postées par des simples internautes sans aucune contrainte ni traitement spécifique de la vidéo. De plus, les événements contenus dans ces vidéos peuvent n'apparaitre qu'à un moment précis de la vidéo. Par exemple, dans les vidéos représentant l'événement « préparer un sandwich », nous pouvons ne voir des scènes illustrant l'événement que sur quelques plans de la vidéo alors que le reste des plans ne représente ni le mouvement ni les objets de l'événement.

Sachant que l'annotation des plans dans les vidéos est une étape manuelle, longue et coûteuse, et qu'une simple réplication du niveau de la vidéo au niveau du plan conduit une mauvaise annotation pour les raisons citées précédemment, nous proposons ici une approche faiblement supervisée qui permettra de produire une annotation au niveau des plans. Cette annotation au niveau des plans est effectuée en se basant sur le contenu des images des plans, un contenu qui peut être visuel, textuel ou audio. La suite de cet article est organisée de la manière suivante : dans la section 2, nous présenterons les travaux connexes à notre travail ; dans la section 3, nous présenterons notre approche en détaillant les différentes étapes menant aux nouvelles annotations au niveau des plans ; enfin dans la section 4 nous présenterons les résultats obtenus lors de l'évaluation de notre approche réalisée dans le cadre de la tâche de détection des événements multimédia (MED) de TRECVID 2011 (Over *et al.,* 2011) .

2. Travaux connexes

Un système de détection automatique d'événements dans les vidéos a pour objectif d'attribuer à une vidéo v un score s représentant la probabilité d'apparition d'un événement e dans v. Ce score se situe donc entre 0 et 1. La détection automatique d'événements dans les vidéos se fait classiquement en trois étapes :

– Extraction de descripteurs à partir des vidéos (signatures) : les descripteurs sont censés représenter le contenu de la vidéo par exemple selon la couleur, la texture, le son, le mouvement ou les points d'intérêt (Lowe, 2004, Laptev, 2005). Ils sont calculés par image, plan ou même vidéo entière ;

– Classification par apprentissage supervisé : l'apprentissage s'effectue sur une collection de vidéos d'apprentissage contenant une quantité équilibrée d'exemples positifs et négatifs de l'événement en question et leurs représentations. Le modèle

d'apprentissage généré serait alors en mesure d'attribuer un score à un couple (v,e) à partir de la signature de la vidéo v. Les SVM sont souvent utilisés ;

– Fusion : Sachant que plusieurs classificateurs ainsi que plusieurs descripteurs peuvent être utilisés, une fusion sera appliquée sur les différents scores obtenus par classificateur et par descripteur. Dans (Ayache *et al.,* 2007, Quénot *et al.,* 2008), des méthodes de fusion sont proposées pour obtenir un seul score global pour un couple (v,e) ;

Ce papier est relié aux travaux concernant la détection automatique d'événements, nous avons utilisé les « sacs de caractéristiques » pour représenter le contenu des plans par des histogrammes de descripteurs et SVM pour l'apprentissage des modèles d'événements. L'annotation automatique des plans à partir d'une annotation des vidéos est similaire à la problématique de dérivation d'étiquettes de pixels à partir d'étiquettes d'images traitée dans (Ries *et al.,* 2012) . Différemment de (Duchenne *et al.,* 2009) nous n'utilisons pas des méthodes de clustering mais des méthodes de sélection par la densité développées dans (Quénot *et al.,* 2012) .

3. Filtrage des plans

Notre objectif est de séparer ce qui est commun aux vidéos représentant un même événement (l'événement lui même) de ce qui est propre à chacune des vidéos (contexte de la vidéo). Pour simplifier, le problème se résume à retrouver des zones de densité entre les plans des différentes vidéos qui contiendront probablement les plans représentant l'événement. Pour former ces zones de densité, comme (Quénot *et al.,* 2012) , nous nous basons sur une intuition que l'on peut formuler des manières suivantes :

– Les plans représentant un événement donné (plans positifs) ont tendance à avoir plus de similarités entre eux qu'avec les plans ne représentant pas un même événement (plans négatifs) ;

– Les représentations des plans positifs ont tendance à être regroupées alors que celles des autres plans se retrouvent souvent éloignées et dispersées ;

– Un plan positif a tendance à avoir une distance moyenne à ses voisins les plus proches moins importante qu'un plan négatif ;

Ces formulations ne sont pas tout à fait équivalentes ni très précises mais correspondent plus ou moins à la même idée. Il n'est pas évident non plus de prouver qu'elle puisse aider à générer des annotations correctes par plan. Le but du travail présenté dans cet article est d'évaluer la justesse de ces hypothèses et la mesure dans laquelle les nouvelles annotations peuvent effectivement améliorer l'apprentissage. Pour la suite, nous détaillerons les étapes nécessaires pour la génération des nouvelles annotations.

3.1. *Matrice de distance*

Soit un descripteur x et une distance d entre descripteurs, nous générons une matrice de distance M de taille $N \times N$ où N est le nombre de plans de l'ensemble des vidéos annotées positives. La matrice M est la matrice de distance entre chaque paire de plans de l'ensemble des plans des vidéos annotées positives. Nous utiliserons simplement pour d la distance euclidienne.

3.2. *Score global*

Une fois que la matrice de distance est générée, nous calculerons un score D intégrant la distance d'un plan par rapport à l'ensemble de ses voisins. Donc chaque plan sera représenté par ce score global. Le choix pour l'expression du score est à déterminer par validation croisée sur l'ensemble de développement. Il existe différentes façons de calculer ce score : par une distance moyenne entre un plan x_i et ses k plus proches voisins, par une distance de médiane entre un plan x_i et ses k plus proches voisins ou une distance pondérée entre un plan x_i et ses k plus proches voisins, de façon à prendre en compte le degré d'éloignement des voisins. Avec $f(k)$ une fonction de pondération, le score sera calculé de la façon suivante :

$$D(x_i) = \frac{\sum_{j=1}^{j=k} f(j) d(x_i, x_{n(i,j)})}{\sum_{j=1}^{j=k} f(j)}$$

$n(i,j)$ étant l'indice du j-ème plus proche voisin de l'échantillon x_i.

3.3. *Choix du seuil*

La dernière étape consiste à fixer un seuil à partir des scores calculés et c'est à partir de ce seuil que nous générerons les nouvelles annotations des plans des vidéos positives. Si le score d'un de ces plans dépasse le seuil fixé, le plan sera considéré en dehors des zones de densité et donc différent des autres plans ; par la suite il sera annoté en « négatif ». Là encore, le seuil peut être calculé de différentes façons : par une moyenne sur les scores, par une médiane sur les scores ou par un pourcentage p de façon à ne garder que $p\%$ des plans positifs les plus proches les uns des autres.

4. Expérimentations

Nous avons évalué notre approche en utilisant une collection de vidéos fournie par TRECVID pour la tâche de détection des événements (MED) 2011. Cette collection contient 45 176 vidéos collectées d'Internet et annotées pour 15 événements complexes au niveau vidéo. La collection est divisée en deux parties, une partie de développement contenant 13 115 vidéos pour environ 370 heures et une partie de test

contenant 32 061 vidéos pour 1200 heures. Dans les événements annotés on trouve : *Changing a vehicle tire, Flash mob gathering, Getting a vehicle unstuck, Grooming an animal, Making a sandwich, Parade, Parkour, Repairing an appliance, etc.*

Sachant que nous ne possédons pas d'annotations au niveau des plans pour la collection utilisée, nous ne pouvons pas évaluer directement notre approche. Nous l'avons évaluée indirectement en mesurant l'impact des nouvelles annotations au niveau du plan sur la détection des événements. Nous avons lancé alors un apprentissage avec les nouvelles annotations générées par notre modèle pour comparer les capacités de détection d'événements de ce nouveau modèle avec celles de la référence (annotations propagées simplement du niveau vidéo au niveau plan). La mesure appropriée pour l'estimation de la qualité de détection de notre système est la moyenne globale sur la précision, la MAP (Mean average precision). Les expérimentations utilisent un descripteur de couleur (histogramme RGB $4 \times 4 \times 4$) et de texture (transformée de Gabor avec 8 orientation et 5 échelles) combinés qui constituent un bon descripteur grâce à sa simplicité et à sa performance dans l'indexation sémantique des vidéos (Delezoide *et al.,* 2011).

Contrairement à nos attentes, nous avons obtenu une amélioration globale relative très faible, de 1.3%. L'analyse des résultats concept par concept a montré que l'amélioration pour certains concepts comme « Birthday party » est bien plus élevée que pour d'autres comme « Parkour » comme on peut l'observer dans les figures 1 et 2. Ceci peut être dû à un biais du système de détection qui se baserait surtout sur le contexte pour détecter les événements. Une comparaison directe des annotations générées au niveau des plans avec une « vérité terrain » pour ces annotations permettra une meilleure évaluation de notre approche.

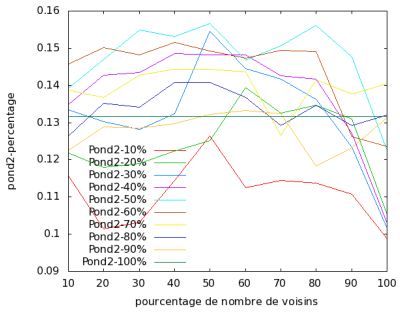

Figure 1. *Les performances pour le concept « Birthday Party »*

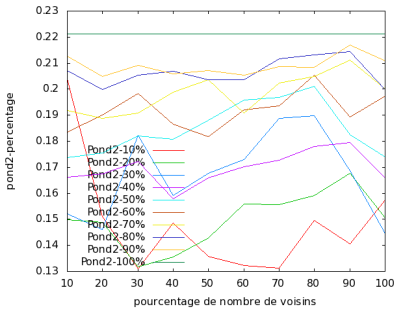

Figure 2. *Les performances pour le concept « Parkour »*

5. Conclusion

Nous avons présenté une méthode permettant de produire des annotations au niveau des plans à partir d'annotations au niveau de vidéos complètes. L'objectif de

cette méthode est de réduire le bruit causé par une simple propagation des annotations du niveau de la vidéo complète vers le niveau plan dans le contexte d'un apprentissage supervisé pour la détection de concepts ou d'événements. Cette méthode utilise le contenu visuel des plans et des vidéos en se basant sur l'idée que les plans contenant un concept ou un événement sont semblables alors que les plans ne représentant pas ce concept ou cet événement sont différents entre eux et du reste des plans. L'approche a été implémentée et évaluée indirectement dans le cadre de la tâche MED de TRECVID 2011 en produisant automatiquement des annotations au niveau des plans et en évaluant ensuite la capacité de détection des événements des modèles entraînés sur ces données. Cette approche n'a pas toujours permis d'améliorer les performances de la détection d'événements mais elle a tout de même apporté des améliorations pour quelques événements. Plusieurs pistes peuvent encore être explorées et testées, l'évaluation directe de notre approche ainsi que l'utilisation d'autres descripteurs plus adaptés aux événements comme les SIFT, SURF, STIP et autres descripteurs basés sur le mouvement.

6. Bibliographie

Ayache S., Quénot G., Gensel J., « Classifier fusion for SVM-based multimedia semantic indexing », *Advances in Information Retrieval*p. 494-504, 2007.

Delezoide B., Precioso F., Gosselin P., Redi M., Merialdo B., Granjon L., Pellerin D., Rombaut M., Jégou H., Vieux R., Bugeau A., Mansencal B., Benois-Pineau J., Boujut H., Ayache S., Safadi B., Thollard F., Quénot G., Bredin H., Cord M., Benoît A., Lambert P., Strat T., Razik J., Paris S., Glotin H., « IRIM at TRECVID 2011 : Semantic Indexing and Instance Search », *TREC Video Retrieval Evaluation workshop*, p. 10 pages, 2011.

Duchenne O., Laptev I., Sivic J., Bach F., Ponce J., « Automatic annotation of human actions in video », *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

Laptev I., « On space-time interest points », *Int. J. Computer Vision*, vol. 64, n° 2, p. 107-123, 2005.

Lowe D. G., « Distinctive Image Features from Scale-Invariant Keypoints », *International Journal of Computer Vision*, vol. 60, n° 2, p. 91-110, 2004.

Over P., Awad G., Fiscus J., F. Smeaton A., Kraaij W., Quénot G., « TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics », *Proceedings of TRECVID 2011*, NIST, USA, dec, 2011.

Quénot G., Benois-Pineau J., Mansencal B., Rossi E., Cord M., Precioso F., Gorisse D., Lambert P., Augereau B., Granjon L. et al., « Rushes summarization by IRIM consortium : redundancy removal and multi-feature fusion », *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, ACM, p. 80-84, 2008.

Quénot G., Thollard F., « Reclassement d'images par le contenu », *CORIA 2012*, Bordeaux, mar, 2012.

Ries C., Lienhart R., « Deriving a discriminative color model for a given object class from weakly labeled training data », *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, p. 44, 2012.