Annotation de vidéos par paires rares de concepts

Abdelkader Hamadi — Philippe Mulhem— Georges Quénot

1. UJF-Grenoble;1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
{Abdelkader.Hamadi, Georges.Quenot, Philippe.Mulhem}@imag.frr

RÉSUMÉ. *La détection d'un concept visuel dans les vidéos est une tâche difficile, spécialement pour les concepts rares ou pour ceux dont il est compliqué de décrire visuellement. Cette question devient encore plus difficile quand on veut détecter une paire de concepts au lieu d'un seul. En effet, plus le nombre de concepts présents dans une scène vidéo est grand, plus cette dernière est complexe visuellement, et donc la difficulté de lui trouver une description spécifique s'accroit encore plus. Deux directions principales peuvent être suivies pour tacler ce problème: 1) détecter chaque concept séparément et combiner ensuite les prédictions de leurs détecteurs correspondants d'une manière similaire à celle utilisée souvent en recherche d'information, ou 2) considérer le couple comme un nouveau concept et générer un classifieur supervisé pour ce nouveau concept en inférant de nouvelles annotations à partir de celles des deux concepts formant la paire. Chacune de ces approches a ses avantages et ses inconvénients. Le problème majeur de la deuxième méthode est la nécessité d'un ensemble de données annotées, surtout pour la classe positive. S'il y a des concepts rares, cette rareté s'accroit encore plus pour les paires formées de leurs combinaisons. D'une autre part, il peut y avoir deux concepts assez fréquents mais il est très rare qu'ils occurrent conjointement dans un même document. Certains travaux de l'état de l'art ont proposé de palier ce problème en récoltant des exemples représentatifs des classes étudiées du web, mais cette tâche reste coûteuse en temps et argent. Nous avons comparé les deux types d'approches sans recourir à des ressources externes. Notre évaluation a été réalisée dans le cadre de la sous-tâche "détection de paire de concepts" de la tâche d'indexation sémantique (SIN) de TRECVID 2013, et les résultats ont révélé que pour le cas des vidéos, si on n'utilise pas de ressources d'information externes, les approches qui fusionnent les résultats des deux détecteurs sont plus performantes, contrairement à ce qui a été montré dans des travaux antérieurs pour le cas des images fixes. La performance des méthodes décrites dépasse celle du meilleur résultat officiel de la campagne d'évaluation précédemment citée, de 9% en termes de gain relatif sur la précision moyenne (MAP).*

CORIA 2014, pp. 239–254, Nancy, 18-21 mars 2014

ABSTRACT. *Single visual concept detection in videos is a hard task, especially for infrequent concepts or for those difficult to model. This question becomes even more difficult in the case of concept pairs. Two main directions may tackle this problem: 1) combine the predictions of their corresponding detectors in a way that is widely used in information retrieval, or 2) build supervised learners for these pairs of concepts by generating annotations based on the occurrences of the two individual concepts. Each of these approaches have advantages and drawbacks. The major problem with the second method is the need of a set of annotated samples, especially for the positive class. If there are some concepts which are infrequent, this scarcity is increasing even more for pairs formed by their combinations. On the other hand, there may be two frequent concepts but they co-occur rarely conjointly in the same document. Some studies suggested to overcome this problem by harvesting samples from the web, but this solution is expensive in terms of time and money. In this work, we compare the two approaches without using any external resources. Our evaluation was carried in the context of the concept pair detection subtask of the TRECVID 2013 semantic indexing (SIN) task, and the results showed that in the case of videos, if we do not use external information resources, the approaches which combine the two concepts detectors can be more efficient than learning based methods, in contrast to what was shown previously in the case of still images. The described methods outperform the best official result of the evaluation campaign cited previously, by 9% in terms of relative improvement on MAP.*

MOTS-CLÉS : *Indexation sémantique, Multimédia, Fusion, Paires de concepts, Concepts visuels, Fusion, multi-SVM TRECVid*

KEYWORDS: *Semantic Indexing, Multimedia, Fusion, Concept pairs, Visual concepts, Fusion, multi-SVM TRECVid*

1. Introduction

Les concepts facilitent l'accès aux sémantiques d'un contenu visuel. Cela rend la détection des concepts visuels une tâche très importante et intéressante dans le domaine de la recherche d'informations multimédia. En plus, rechercher la co-occurrence d'un ensemble de concepts visuels dans des images/vidéos non annotées est une étape importante pour répondre à des requêtes complexes des utilisateurs. En effet, ces requêtes sont généralement exprimées via un ensemble de termes sémantiques. D'une autre part, considérer qu'un document multimédia est indexé par plusieurs concepts est utile : une simple combinaison d'un ensemble de concepts peut représenter d'autres sémantiques pouvant être complexes. Par exemple, la combinaison des concepts "neige", "montagne" et "personne(s) entrain de se déplacer" pourrait être liée au fait qu'on est entrain de voir une scène de "skieur" ou "une compétition de ski". Si la détection d'un seul concept est une dure tâche, spécialement pour ceux qui sont rares ou dont il est difficile de représenter ou décrire visuellement d'une manière efficace, cette difficulté s'accroit encore plus quand on veut vérifier l'occurrence conjointe de N concepts dans un même document. De plus, une scène même avec juste deux concepts tend à être complexe visuellement, et le défi reste difficile même dans le cadre d'une paire de concepts. Cette remarque est confirmée par les résultats médiocres en termes de performance, obtenus par les travaux qui se sont focalisés auparavant sur la détection de paire de concepts dans les images/vidéos. D'une autre part, les approches de l'état de l'art concernant la détection de concepts dans les vidéos sont majoritairement basées sur l'apprentissage supervisé. Ces méthodes nécessitent des corpus de données dont l'obtention s'avère coûteuse en temps et argent. Donc, si on veut construire un modèle spécifique pour chaque paire de concepts, on se heurtera forcément au problème de nécessité de données annotées. Il est évident que le nombre d'échantillons positifs pour une paire de concepts dans un corpus est beaucoup plus petit que le nombre d'exemples positifs pour chacun des deux concepts formant la paire. Notons aussi que s'il y a des concepts rares, cette rareté s'accroit encore plus pour les paires formées de leurs combinaisons. D'une autre part, il peut y avoir deux concepts assez fréquents mais il est très rare qu'ils occurrent conjointement dans un même document. Certains travaux de recherche de l'état de l'art ont proposé des méthodes qui font l'effort de récolter des exemples positifs et négatifs du web social, mais cette solution reste coûteuse en termes de temps et d'argent, surtout quand on ne dispose pas de moyens nécessaires pour les mettre en œuvre. Ces remarques restent valides et la situation sera encore plus compliquée si on considère plus de deux concepts à la fois. Il est donc justifié de s'attaquer au problème de la détection de paires de concepts avant de passer à la détection de plus de deux concepts. D'une autre part, si une paire de concepts est fréquente dans un corpus, cela implique que les deux concepts la formant sont beaucoup plus fréquents. De ce fait, l'intersection des résultats des détecteurs des deux concepts ou la génération d'un modèle spécifique pour la paire, pourraient donner des résultats plus ou moins satisfaisants, grâce à la disponibilité des données. Le grand défi concerne le cas des paires rares dans lequel un classifieur supervisé aura sûrement du mal à trouver une séparation efficace à cause du manque de données. À base de ce raisonnement et détails, nous nous focalisons dans

ce travail exclusivement sur la détection de *paires* (bi-concept) *rares* de concepts dans les *vidéos*.

Une alternative à l'idée de générer un modèle spécifique pour chaque bi-concept consiste en la combinaison des scores de prédiction des détecteurs des deux concepts formant la paire. En effet, rechercher une paire de concepts (A,B) est similaire à la recherche de l'occurrence de A ET B dans le même document. On peut penser au modèle booléen pondéré de Fox (1983) ou a des implications de logique floue (Lukasiewicz, Zadeh). Un des avantages de cette idée c'est qu'elle ne nécessite pas de traitements additionnels coûteux car le problème consiste en gros à fusionner les réponses relatives aux détecteurs des deux concepts. Cependant, ces approches ne peuvent pas prendre en considération les spécificités visuelles qui caractérisent les co-occurrences de concepts dans un même plan vidéo. Par exemple, les avions pourraient co-occurer avec les bâtiments et dans ce cas, les avions pourraient avoir des spécificités visibles, comme de visibles équipements d'atterrissage. Ce genre de spécificités n'est pas pris en compte en fusionnant les scores de détection. Laquelle des deux types d'approches faut-il utiliser ? Le but de ce travail est de répondre à cette question, tout en gardant à l'esprit que nous ne voulons pas une solution imposant des traitements très coûteux.

Dans l'état de l'art, les méthodes qui ont adressé le problème de la détection de paires de concepts n'ont pas été très convaincantes en termes de performance. Dans TRECVID 2012, une sous-tâche de la tâche d'indexation sémantique (SIN) a été introduite pour la première fois, pour traiter le problème de la détection de paires de concepts, à laquelle six équipes ont participé et le meilleur système a obtenu une valeur de précision moyenne (MAP) qui n'a pas dépassé pas les 8%, sachant que la performance des meilleurs systèmes pour détecter un seul concept dans la tâche SIN de TRECVID 2012 a été de l'ordre de 30% en MAP. Cela montre l'étendu du défi concernant la problématique abordée.

Le reste de l'article est organisé comme suit. La section 2 présente un ensemble de travaux ayant un lien avec la problématique étudiée. Dans la section 3 nous décrivons une approche permettant de générer un modèle spécifique pour chaque paire de concepts et quelques variantes de la méthode qui consiste à fusionner les sorties de détecteurs de concepts individuels. Les expérimentations sont décrites et discutées dans la section 4. Nous concluons dans la section 5 en citant des perspectives pour les travaux à venir.

2. État de l'art

Pour détecter une paire de concept, on pourrait combiner les résultats des détecteurs des concepts formant la paire. Même si la performance des détecteurs individuels est raisonnablement bonne, la fusion de leurs scores donne de mauvais résultats (Li *et al.*, 2012). Un nombre considérable de travaux de recherche se sont facalisés sur cet axe. Dans (Aly *et al.*, 2008), les auteurs utilisent des règles de produit comme fonction de fusion, et plusieurs travaux (Wei *et al.*, 2011 ; Snoek

et al., 2007 ; Li *et al.*, 2007 ; Chang *et al.*, 2006 ; Yan et Hauptmann, 2003) ont étudié des fusion linéaires. Ces travaux ne concernent pas spécifiquement la détection des paires de concepts (bi-concepts).

En ce qui concerne les études sur la détection de bi-concepts dans les documents multimédias, à savoir des images fixes, (Li *et al.*, 2012) proposent de générer un modèle par bi-concept à partir des échantillons collectés du web social. Les auteurs sont arrivés à la conclusion qu'apprendre des modèles de bi-concepts est mieux qu'une fusion linéaire de détecteurs de concepts individuels. La question reste cependant ouverte pour le cas des vidéos. Dans (Wang et Forsyth, 2009), les auteurs présentent une méthode pour apprendre des attributs visuels (e.g., red, metal, spotted) et des classes d'objets (e.g., car, dress, umbrella) ensemble, mais ce ne sont réellement pas des bi-concepts parce que le couple refère à un et un seul et même concept (e.g., Le couple (voiture, rouge) fait référence à une même chose qui est une voiture).

Nous avons remarqué qu'il y a très peu d'études dans l'état de l'art qui ont adressé d'une manière spécifique le problème de la détection de paires de concepts dans les documents multimédia, encore moins pour le cas des vidéos. Nous rappelons que dans le cadre de ce travail, nous utilisons certaines approches existantes dans le domaine de la recherche d'information pour les appliquer et les comparer dans le cadre de la détection de *paires rares* de concepts dans les *vidéos*. Le choix est fait de manière à garder une certaine généralité et un peu de réalisme, en évitant de créer explicitement des traitements très coûteux. Nous avons choisi des méthodes simples et faciles à mettre en œuvre, et surtout ne nécessitant pas de ressources externes d'information.

Le travail que nous proposons ici se heurte au problème de la rareté des échantillons positifs. Plusieurs chercheurs se sont focalisé sur ce problème problème. L'apprentissage actif demeure une solution efficace. Pour pallier le problème de la rareté des échantillons ou plus précisément le problème des classes déséquilibrées, des chercheurs ont proposé de re-échantillonner les données de développement initiales, soit par un sur-échantillonnage de la classe minoritaire, ou par un sous-échantillonnage de la classe majoritaire. Ceci est fait de telle manière que les classes soient présentées d'une manière plus équilibrée (Bishop, 2007 ; Chawla *et al.*, 2002 ; Weiss et Provost, 2001). Le sur-échantillonnage induit à un temps d'apprentissage plus long et au problème de gourmandise en mémoire, en raison de l'augmentation du nombre d'instances d'entrainement. En effet, le coût de calcul sera plus élevé pour pré-traiter les données. D'une autre part, le sous-échantillonnage peut conduire à une perte d'information, en raison du fait qu'il peut ignorer des données utiles qui pourraient être importantes pour le processus d'apprentissage. Certains autres chercheurs ont pensé à entrainer leurs classifieurs sur des données auxiliaires ou additionnelles (Li *et al.*, 2012 ; Wu et Dietterich, 2004) obtenues depuis une source externe. Il existe également une autre solution, qui consiste en l'utilisation des échantillons d'une seule classe, typiquement la classe cible (positive), au lieu d'opter pour une classification binaire (les classes positive et négative), comme SVM à une seule classe (SVM one-class) (Chen *et al.*, 2001), mais la performance de ce genre de méthodes dépend de la représentativité des échantillons d'apprentissage.

Nous considérons par la suite, deux catégories d'approches pour la détection de paires rares de concepts dans les vidéos. La première construit un modèle de bi-concept en utilisant une méthode d'apprentissage d'ensembles (Ensemble learning), avec une philosophie similaire à celle de (Li *et al.*, 2012), en considérant quelques adaptations aux : 1) paires de concepts, 2) cas des vidéos, et 3) méthodes d'apprentissage de l'état de l'art pour l'indexation des vidéos. Nous avons choisi ce type de classifieur pour traiter le problème de la rareté des échantillons positifs. Dans le deuxième type d'approches, les détecteurs de concepts individuels sont fusionnés en utilisant de schémas simples, sans aucune étape d'apprentissage.

3. Propositions

3.1. *Modèles de paire de concepts*

Nous proposons d'utiliser une méthode qui consiste à générer un modèle spécifique pour chaque paire de concepts. Contrairement à (Li *et al.*, 2012), nous ne considérons pas de données supplémentaires, mais nous nous servons uniquement des données relatives aux concepts singuliers.

Nous modélisons cette approche par un 10-tuple :

$< C^{single}, C^{paire}, E, E_D, E_T, F_{descr}, F_{ann}^{single}, F_{ann}^{paire}, F_{detect}^{paire}, F_{d-fuse} >$

Où :

– C^{single} : L'ensemble de concepts singuliers ;

– C^{paire} : L'ensemble des paires de concepts. Pour chaque paire (A, B), on a nécessairement $A \in C^{single}$ et $B \in C^{single}$;

– E : L'ensemble des échantillons ;

– E_D : $E_D \subset E$; L'ensemble d'apprentissage annoté manuellement par au moins une concept $c \in C^{single}$;

– E_T : $E_T \subset E$; L'ensemble des échantillons de test ;

– F_{ann}^{single} : $E_D \times C^{single} \rightarrow \{0, 1, -1\}$; Cette fonction renvoie un entier positif correspondant à l'annotation (dans notre cas manuelle) d'un échantillon $e \in E_D$ par un concept $c \in C^{single}$. La valeur renvoyée appartient à l'ensemble $\{0, 1, -1\}$ où 0 : signifie que l'exemple est annoté négativement, 1 : signifie que l'exemple est annoté positif et la valeur -1 est renvoyée quand l'échantillon n'est pas annoté ;

– F_{descr} : $E \rightarrow R^{dim}$; cette fonction donne une description d'un échantillon $e \in E$. dim est la taille du descripteur extrait ;

– F_{ann}^{paire} : $E_D \times C^{paire} \rightarrow \{0, 1, -1\}$; cette fonction renvoie une valeur entière correspondant à l'annotation d'un échantillon $e \in E_D$ par une paire de concept $pc \in C^{paire}$. La valeur renvoyée appartient à l'ensemble {0, 1, -1} où 0 : signifie que l'exemple est annoté négativement, 1 : signifie que l'exemple est annoté positif et la valeur -1 est renvoyée quand l'échantillon n'est pas annoté ;

– $F_{detect}^{paire} : E \times C^{paire} \rightarrow R$. Cette fonction détecte une paire de concepts dans un échantillon $e \in E$ et renvoie une valeur correspondant au score de détection ;

– F_{d-fuse} : Cette fonction renvoie une valeur correspondant à la fusion d'un certains nombre de scores. Comme par exemple la fusion des scores obtenus en utilisant différents descripteurs.

Pour instancier ce modèle nous définissons dans ce qui suit les différentes fonctions. Nous nous basons pour cela uniquement sur les données relatives aux concepts singuliers.

F_{ann}^{single} est remplacée par le rôle d'un expert humain chargé d'annoter manuellement l'ensemble de données d'apprentissage. Étant donnée une paire de concept (c_1,c_2), nous définissons la fonction F_{ann}^{paire} comme l'intersection des annotations par c_1 et c_2 :

$$F_{ann}^{paire}(e,(c_1,c_2)) = \begin{cases} 1\ (positif)\ si\ F_{ann}^{single}(e,c_1) = 1\ et\ F_{ann}^{single}(e,c_2) = 1 \\ 0\ (négatif)\ si\ F_{ann}^{single}(e,c_1) = 0\ ou\ F_{ann}^{single}(e,c_2) = 0 \\ -1\ (non\ annoté)\ sinon \end{cases}$$

Généralement il y a peu d'exemples positifs et beaucoup d'exemples non annotés. Par conséquent, la fonction F_{ann}^{paire} donne une matrice d'annotations creuse. Autrement dit, F_{ann}^{paire} donnerait très peu d'exemples positifs et beaucoup d'exemples non annotés. Il faut savoir aussi que les échantillons non annotés ne seront pas utilisés dans l'étape d'apprentissage. Ce phénomène complique la situation et affecte négativement la performance des classifieurs. Pour pallier à cet inconvénient, nous proposons d'utiliser comme fonction F_{detect}^{paire}, une méthode d'ensemble pour l'apprentissage et plus précisément le "Bagging". Ce genre d'approche est compatible avec le problème de classes déséquilibrées. Nous choisissons la méthode décrite dans (Safadi et Quénot, 2010), pour ses bons résultats dans le contexte de la détection de concepts visuels dans les vidéos et sa compatibilité avec le problème de classes déséquilibrées. Elle consiste à combiner m classifieurs via une stratégie de "Bagging" où chacun d'entre eux utilise tous les échantillons d'apprentissage de la classe dominée (dans notre cas : la classe positive), et un ensemble d'échantillons de la classe dominante (dans notre cas : la classe négative) est tiré aléatoirement avec remise (boostrap), avec :

$$m = (f_{neg} * N_{neg})/(f_{pos} * N_{pos})$$

où N_{pos} : est le nombre d'exemples positifs, N_{neg} : est le nombre d'échantillons négatifs, f_{neg} et f_{pos} sont des paramètres (entiers positifs non nuls) relatifs aux classes positive et négative, respectivement. Nous rappelons à ce stade, que l'annotation concerne une paire de concepts et non un concept individuel. f_{pos} gère la proportion des échantillons de la classe dominante qu'on veut utiliser, par rapport au nombre d'exemples de la classe dominée (e.g. deux fois plus d'exemples négatifs que positifs). f_{neg} quant à lui, permet de contrôler à l'aide de f_{pos} le nombre de classifieurs souhaité. L'ensemble E_D est dévisé en m sous-ensembles, où chaque sous-ensemble contient tous les exemples positifs contenus dans E_D et ($f_{pos} * N_{pos}$) exemples négatifs sont

tirés aléatoirement avec remise. Ensuite chacun des m classifieurs est entrainé sur un sous-ensemble différent. On remarque que la contrainte $f_{neg} * N_{neg} \geq f_{pos} * N_{pos}$ doit être vérifiée. Finalement les scores des m classifieurs sont fusionnés en utilisant n'importe quelle fonction possible. Dans notre cas, nous choisissons une moyenne simple des scores de prédictions, pour sa simplicité d'une part et aussi pour ses bons résultats dans le contexte abordé et les données utilisées dans nos expérimentations. Plus la valeur de m est grande, meilleure est la performance finale. Pour le type des classifieurs combinés, nous retenons dans notre travail :"SVM", pour ses bons résultats dans le domaine de l'indexation des vidéos. Nous appellerons par la suite, cette méthode globale de classification : "multi-SVMs".

Pour décrire un plan vidéo (F_{descr}), nous proposons d'utiliser plusieurs types de descripteurs et ce, pour chaque plan vidéo. F_{descr} correspond donc à la méthode d'extraction de descripteurs de bas niveau (e.g., visuel, audio, etc) des plans vidéo. Nous aborderons les détails concernant les descripteurs utilisés dans la section4. Pour la fusion tardive (F_{d-fuse}) des résultats obtenus par les différents descripteurs utilisés, nous choisissons d'utiliser une somme pondérée où les poids sont des scores de confiance :

$$F_{d-fuse}(sc_1, sc_2, \ldots, sc_N) = \frac{1}{N} \sum_{i=1}^{N} sc_i * Conf[i]$$

où sc_i est le score obtenu par le i^{eme} résultat (en utilisant le i^{eme} descripteur), $Conf[i]$: est un score de confiance concernant le i^{eme} résultat (i^{eme} descripteur), et N est le nombre de résultats à fusionner. Nous choisissons la précision moyenne calculée sur un corpus de développement comme score de confiance. Le score renvoyé par F_{d-fuse} sera le score final de détection de la paire dans l'échantillon en question.

Nous appellerons dans ce qui suit, le résultat obtenu en utilisant cette méthode : $learnDouble$.

3.2. *Fusion de détecteurs de concepts individuels*

Les méthodes qui consistent à fusionner les scores de détection ont un grand avantage sur la méthode d'apprentissage directe : elles ne nécessitent pas de lourds traitements pour les nouvelles requêtes. Les méthodes que nous allons utiliser n'ont aucun paramètre à apprendre et/ou à optimiser, ce qui permet leur application directement et facilement sur n'importe quelle paire de concepts. Nous proposons le modèle général suivant :

$< C^{single}, C^{paire}, E, E_D, E_T, F_{descr}, F^{single}_{ann}, F^{c}_{detect}, F^{paire}_{detect} >$ Où :

– C^{single}, C^{paire}, E, $E_D \subset E$, $E_T \subset E$, F^{single}_{ann}, F_{descr}, sont définis dans le modèle de la section 3.1.

– $F^{c}_{detect} : E \rightarrow R$. Cette fonction détecte un concept $c \in C^{single}$ dans un échantillon $e \in E$ et renvoie une valeur correspondant au score de détection. Elle

peut être différente pour chaque concept individuel c, comme elle peut être la même fonction pour tous les concepts dans une approche générique ;

– $F^{paire}_{detect} : E \times C^{paire} \rightarrow R$. Cette fonction détecte une paire de concept $pc \in C^{paire}$ dans un échantillon $e \in E$ et renvoie une valeur correspondant au score de détection ;

Pour instantier le modèle on se contentera de définir F^{c}_{detect} et F^{paire}_{detect}, puisque les paramètres F_{descr}, F^{single}_{ann} ne changeront pas par rapport au modèle décrit dans la sous-section précédente. Nous choisissons comme fonction F^{c}_{detect} la même méthode choisie pour la génération de modèle pour les paires de concepts : "multi-SVMs" (Safadi et Quénot, 2010), dont nous avons résumée le principe de fonctionnement précédemment. Ce choix est justifié d'une part par la compatibilité de la méthode avec le problème des classes déséquilibrées et donc la rareté des exemples positifs, et d'une autre part pour ses bons résultats. La seule différence c'est qu'il faut utiliser comme annotations le résultat de la fonction F^{single}_{ann}, qui dans notre cas est remplacée par un expert humain fournissant des annotations manuelles.

En ce qui concerne la fonction F^{paire}_{detect}, nous allons examiner plusieurs méthodes de fusion de scores. Nous avons fait dans nos expériences le maximum pour rendre ou garder ces scores de probabilités homogènes. Cela peut être fait en considérant la normalisation de Platt (Platt, 2000) des sorties des classifieurs, et/ou en utilisant une fonction de fusion appropriée. En effet, cette normalisation a un effet sur l'efficacité des méthodes envisagées. Nous proposons d'utiliser les méthodes de fusion suivantes :

– Une fusion linéaire, que nous appellerons *linFus* : moyenne arithmétique des scores. Ce type de fusion est listé dans (Li *et al.*, 2012) :

$$F^{paire}_{detect}(e,(c_1,c_2)) = (F^{c_1}_{detect}(e) + F^{c_2}_{detect}(e))/2$$

– Méthode basée sur la notion de probabilité, appelée dans ce qui suit *prodFus*, ou moyenne géométrique des scores. Cette méthode considère que les scores sont des probabilités et que ces probabilités sont obtenues par des détecteurs conditionnellement indépendants (la racine carrée ne change pas l'ordre des shots) :

$$F^{paire}_{detect}(e,(c_1,c_2)) = \sqrt{F^{c_1}_{detect}(e).F^{c_2}_{detect}(e)}$$

– Une fusion basée sur les rangs que nous appellerons *rankFus*. Cette méthode combine les résultats de détection en se basant sur la sommation des positions des rangs des documents à partir de deux listes triées de résultats. Elle est similaire à *linFus*, à l'exception que les scores sont remplacés par les rangs qui sont eux mêmes dérivés des scores et normalisés entre 0 et 1 (Smeaton, 1998). Les rangs ne sont pas homogènes aux probabilités, puisque le passage des scores aux probabilités induit une perte d'informations. Le calcul du score final de détection passe par les étapes suivantes :

1) Pour chaque paire $(c_1, c_2) \in C^{paire}$ convertir les scores de détection de c_1 en scores représentant le rang de chaque échantillon parmi tous les autres :

$$sc^{rang}(e, c_1) = \frac{N - F_{rank}(F^{c_1}_{detect}(e)) + 1}{N}$$

où F_{rank} est une fonction qui renvoie le rang du document e en comparant son score de détection $F^{c_1}_{detect}(e)$ avec ceux de tous les autres, et N est le nombre des échantillons test ;

2) Trier les éléments test selon les scores $sc^{rang}_{c_1}$, et les deviser en nb bins. La valeur de nb peut être optimisée par validation croisée ou fixée dès le départ.

3) Dans chaque bin, re-trier ses éléments en fonction de $F^{c_2}_{detect}(e)$ et calculer de nouveaux scores ($score^{rang}_{c_1}$) basés sur les rangs comme dans l'étape 1 par exemple ;

4) Reprendre les étapes 1, 2 et 3 en inversant les rôles des concepts c_1 et c_2

5) Combiner les scores $score^{rang}_{c_1}$ et $score^{rang}_{c_2}$. Par exemple, comme suit :

$$F^{paire}_{detect}(e, (c_1, c_2)) = \frac{1}{2}(score^{rang}_{c_1}(e) + score^{rang}_{c_2}(e))$$

– Une approche inspirée de l'approche booléenne étendue de (Salton *et al.*, 1983), que nous nommerons *boolFus*, qui considère une paire de concepts comme la conjonction de deux concepts :

$$F^{paire}_{detect}(e, (c_1, c_2)) = 1 - \sqrt{((1 - F^{c_1}_{detect}(e))^2 + (1 - F^{c_2}_{detect}(e))^2)/2}$$

Le score renvoyé par $F^{paire}_{detect}(e, (c_1, c_2))$ sera le score final de détection de la paire (c_1, c_2) dans le plan e.

4. Expérimentations et résultats

4.1. *Données*

Nous avons évalué les approches décrites dans les sections précédentes dans le cadre de la sous-tâche "détection de paire de concepts" de la tâche d'indexation sémantique de TRECVID 2013. Les annotations pour les concepts singuliers ont été fournies par l'annotation collaborative de TRECVID (Ayache et Quénot, 2008). Cette sous-tâche définit 10 paires de concepts rares et la performance est mesurée par la précision moyenne inférée (InfAP).

Nous avons généré les annotations par paire de concepts à partir de celles des concepts singuliers, comme décrit dans la section3.1 (via la fonction F^{paire}_{ann}). Le nombre d'exemples positifs résultant est très petit, spécialement pour certains bi-concepts. Le tableau1 montre les paires de concepts utilisées ainsi que des détails

sur leurs fréquences dans le corpus d'apprentissage utilisé. Nous pouvons voir que quatre parmi les dix paires de concepts ont moins de 10 exemples positifs. Nous notons aussi que pour faire la validation croisée, ce corpus d'apprentissage doit être divisé en sous-parties, ce qui diminuera encore plus le nombre d'échantillons à l'entrée des classifieurs. Cela reflète la difficulté de la tâche.

Tableau 1. *Fréquences des paires de concepts dans le corpus d'apprentissage TREC-VID 2013. Les annotations sont le résultat d'une intersection des annotations des deux concepts singuliers*

Paires de concepts	# positifs	# negatifs	#non annotés
Telephones+Girl	1	18918	527004
Kitchen+Boy	12	43845	502066
Flags+Boat_Ship	4	10123	535796
Boat_Ship+Bridges	27	39200	506696
Quadruped+Hand	26	18392	527505
Motorcycle+Bus	9	82490	463424
Chair+George_Bush	41	23881	522001
Flowers+Animal	67	41875	503981
Explosion_Fire+Dancing	0	19550	526373
Government_Leader+Flags	321	12828	532774

Pour la description du contenu des plans vidéo, nous avons utilisé au total de 66 variantes de 12 types de descripteurs fournis par le groupe IRIM (Ballas *et al.*, 2012). Les variantes d'un même descripteur se différencient par de légers détails comme par exemple, la taille du dictionnaire utilisé pour la génération des sacs de mots. Nous avons considéré des descripteurs de type : transformation Gabor pour la texture, histogrammes de couleurs, SIFT, STIP, VLAT, Percepts...Tous ces descripteurs sont répertoriés et décrits dans (Ballas *et al.*, 2012).

Pour la détection des concepts singuliers, nous avons utilisé un classificateur supervisé de type multi-SVMs. Comme décrit précédemment, Multi-SVMs consiste à combiner un nombre de classificateurs de type SVM dans un schéma de "Bagging", en prenant pour chacun d'eux tous les échantillons de la classe positive et en sélectionnant un certain nombre d'échantillons de la classe négative via un tirage aléatoire avec remise (boostrap). Comme entrée aux Multi-SVMs, les descripteurs cités ci-dessus sont extraits de chaque plan vidéo. Quant à leurs sorties, elles sont finalement fusionnées pour chaque plan vidéo afin de calculer un score de classification. Cela est réalisé via une moyenne simple. Pour chaque plan vidéo, les scores obtenus par les différents descripteurs sont fusionnés en utilisant une méthode de fusion hiérarchique décrite dans (Tiberius Strat *et al.*, 2012), pour donner un score final de détection d'un concept singulier dans le plan vidéo concerné. Finalement les différentes méthodes décrites dans la section3.2 sont appliquées en se basant sur les scores finaux précédemment calculés.

Pour l'approche consistant à générer un modèle par paire de concepts, les détecteurs sont construits en suivant globalement le même schéma de classification que pour les détecteurs de concepts singuliers (multi-SVMs comme classifieur, même descripteurs). La différence se résume sur les annotations utilisées. Finalement les résultats obtenus par les différents descripteurs sont fusionnés en utilisant la fonction F_{d-fuse} décrite dans la section3.1, qui est une moyenne pondérée des sorties des Multi-SVMs où les poids sont les valeurs de la précision moyenne calculées sur un corpus de développement. Nous notons qu'à ce stade, nous avons privilégié la somme pondérée sur la fusion hiérarchique (Tiberius Strat *et al.*, 2012), parce que nous avons constaté que cette dernière ne marche pas bien pour le cas des modèles par paire de concepts, même si ses résultats dans le cas des concepts singuliers sont assez bons.

4.2. *Résultats*

Tableau 2. *Résultats (InfAP) sur le corpus de test : InfAP pour les 10 paires de concepts évaluées dans TRECVID 2013. Les paramètres d'apprentissage concernant l'approche "learnDouble" sont optimisés par validation croisée.*

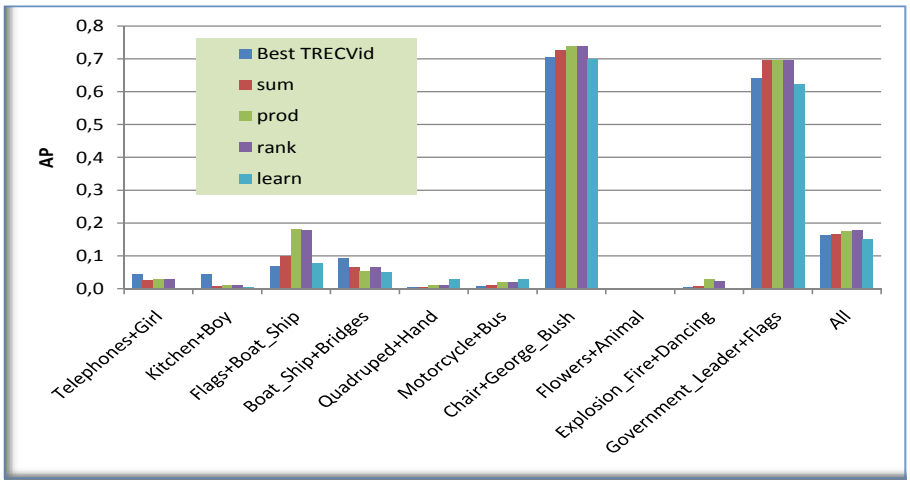
Système	MAP
Meilleure soumission TRECVid 2013	0.1616
linFus	0.1613
prodFus	0.1761
rankFus	0.1767
boolFus	0.1724
learnDouble	0.1514

Le tableau 2 présente les résultats obtenus en termes de précision moyenne inférée InfAP pour les approches considérées. Dans ce tableau, la première ligne présente le meilleur système officiellement évalué dans la sous-tâche "détection de paires de concepts" de TRECVID 2013 . Le groupe des quatre lignes suivantes concernent les résultats obtenus par les méthodes de fusion et la dernière ligne présente le résultat obtenu par la méthode d'apprentissage. Les approches de fusion sont nettement plus performantes que celle d'apprentissage . Parmi les méthodes de fusion, les meilleures sont "prodFus" et "rankFus". Nous notons dans ce contexte, que dans (Li *et al.*, 2012), les auteurs sont arrivés à une autre conclusion concernant la détection de paires de concepts dans les images fixes. En effet, les auteurs ont conclu que la génération d'un modèle spécifique par paire concept est plus efficace que la combinaison de détecteurs de concepts individuels. Or, leurs conditions d'expérimentations diffèrent des nôtres concernant plusieurs points. Premièrement, le type de données considéré dans les deux travaux est différent (vidéos dans notre travail vs. les images fixes dans (Li *et al.*, 2012)). D'une autre part, en plus du fait que dans (Li *et al.*, 2012), une collecte de bons exemples positifs et négatifs précède la phase d'apprentissage, dans notre travail, nous considérons des paires rares de concepts. Nous rappelons ici, que quatre paires de concepts ont moins de 10 échantillons positifs pour l'apprentissage (1

pour Telephones+Girl, 4 pour Flags+Boat_Ship, 9 pour Motorcycle+Bus et 0 pour Explosion_Fire+Dancing), expliquant pourquoi la performance est inférieure pour l'approche "learnDouble". Nous notons aussi, que pour certaines paires de concepts pour lesquelles il y a trop peu d'exemples positifs, il s'est avéré impossible de dérouler la méthode d'apprentissage, ce qui a conduit à des résultats médiocres, comme c'est le cas pour les deux paires "Telephones+Girl " et "Explosion_Fire+Dancing" (voir la figure 1). À l'opposé de notre situation, la fréquence moyenne des paires de concepts utilisées dans (Li *et al.*, 2012) est d'environ 3900. D'une autre part, la méthode utilisée pour générer les détecteurs de concepts individuels a certainement aussi un impact sur le résultat de la fusion. En effet, chaque méthode capture des informations sémantiques différentes. Dans (Li *et al.*, 2012), les auteurs disent que la recherche de bi-concepts exige des détecteurs individuels avec une bonne capacité de discrimination, ce qui est le cas dans notre travail. En effet, notre système de détection des concepts individuels a été classé deuxième dans la tâche d'indexation sémantique (SIN) de la compagne d'évaluation TRECVID 2013.

On remarque aussi dans le même tableau que les résultats obtenus dépassent le meilleur résultat officiel de TRECVID 2013, par 9,34 % (100 * (0,1767 -0,1616) / 0,1616 = 9,34 %). Nous pouvons voir à l'aide du test "Student" apparié bilatéral que, même si les valeurs de MAP diffèrent beaucoup, les résultats du tableau 2 ne sont pas statistiquement différents, avec une valeur de $p < 5$ % , cela est dû principalement au nombre petit des paires de concepts considérées (uniquement 10 paires). Les méthodes présentées ici ne sont pas toutes incluses dans nos soumissions officielles à la sous-tâche de détection de paires de concepts TRECVID 2013, cela est dû au fait que les résultats définitifs n'étaient pas achevés avant la date limite. Malgré ça, nous avons quand même obtenu la quatrième place.

Figure 1. *Résultats des différentes approches pour la détection de paires de concepts sur la collection TRECVID 2013, et comparaison avec le meilleur résultat de la tâche SIN de TRECVID 2013.*



La figure Fig.1 décrit les résultats par paire de concepts en termes de précision moyenne (AP) obtenus par les différentes approches considérées et une comparaison avec le meilleur système officiellement évalué dans TRECVID 2013. Nous pouvons voir que l'approche fondée sur l'apprentissage d'un modèle par paire de concepts est la plus efficace pour les paires : "Quadruped+Hand" "et "Motorcycle+Bus", mais cette performance ne compense pas l'échec sur les autres paires. Nous constatons aussi que dans la valeur de MAP globale, il y a les valeurs de AP de deux paires qui dominent : "Chair+George_Bush" et "Government_Leader+Flags", et que les différentes méthodes sont assez mauvaises en ce qui concerne la détection de certaines paires de concepts. Cela nous motive à utiliser à l'avenir des ressources externes d'informations pour pallier cet inconvénient.

5. Conclusion et perspectives

Nous avons comparé deux types d'approches pour détecter des paires (bi-concepts) rares de concepts visuels dans les vidéos. La première consiste en la construction d'un modèle par bi-concept en générant des annotations en se basant sur celles des concepts singuliers, et en utilisant une méthode d'apprentissage d'ensembles. La deuxième est basée sur la combinaison des sorties de détecteurs de concepts formant la paire, et ne nécessite aucune étape supplémentaire d'apprentissage. Contrairement à ce qui a

été observé dans des travaux antérieurs concernant les images, pour le cas des vidéos, mais sans recourir à des ressources externes, l'approche de fusion des deux détecteurs donne de meilleurs résultats par rapport à la méthode basée sur l'apprentissage direct d'un modèle par paire de concepts. De plus, les approches de fusion sont également beaucoup plus faciles à mettre en œuvre et ne nécessitent ni un re-apprentissage ni de réglages et optimisations des paramètres.

Pour les travaux à venir, nous considérons l'intégration des informations externes (e.g., le contexte, des ontologies, des annotations supplémentaires, récolte des échantillons positifs,...) dans nos approches. Nous pensons que cela pourrait améliorer encore plus les résultats. Par ailleurs, nous sommes également intéressés par la considération de la détection de plusieurs concepts au lieu de seulement deux. Un autre objectif intéressant serait d'appliquer nos méthodes sur les mêmes corpus utilisés dans l'état de l'art pour confirmer ou infirmer certaines hypothèses, surtout celles qui sont en conflit avec les résultats obtenus dans ce travail.

6. Bibliographie

Aly R., Hiemstra D., de Vries A., de Jong F., « A Probabilistic Ranking Framework using Unobservable Binary Events for Video Search », *7th ACM International Conference on Content-based Image and Video Retrieval, CIVR 2008*, ACM, New York, NY, USA, p. 349-358, July, 2008.

Ayache S., Quénot G., « Video corpus annotation using active learning », *Proceedings of the IR research*, ECIR'08, Springer-Verlag, Berlin, Heidelberg, p. 187-198, 2008.

Ballas N., Labbé B., Shabou A., Le Borgne H., Gosselin P., Redi M., Merialdo B., Jégou H., Delhumeau J., Vieux R., Mansencal B., Benois-Pineau J., Ayache S., Hamadi A., Safadi B., Thollard F., Derbas N., Quénot G., Bredin H., Cord M., Gao B., Zhu C., tang Y., Dellandrea E., Bichot C.-E., Chen L., Benot A., Lambert P., Strat T., Razik J., Paris S., Glotin H., Ngoc Trung T., Petrovska Delacrétaz D., Chollet G., Stoian A., Crucianu M., « IRIM at TRECVID 2012 : Semantic Indexing and Instance Search », *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, 2012.

Bishop C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edn, Springer, October, 2007.

Chang S.-F., Hsu W., Jiang W., Kennedy L., Xu D., Yanagawa A., Zavesky E., « Columbia university TRECVID-2006 video search and high-level feature extraction. in Proc. TRECVID Workshop », *In Proc. TRECVID Workshop*, 2006.

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., « SMOTE : Synthetic Minority Over-sampling Technique », *Journal of Artificial Intelligence Research*, vol. 16, p. 321-357, 2002.

Chen Y., Zhou X., Huang T. S., « One-class svm for learning in image retrieval », p. 34-37, 2001.

Li X., Snoek C. G. M., Worring M., Smeulders A., « Harvesting Social Images for Bi-Concept Search », *Multimedia, IEEE Transactions on*, vol. 14, n° 4, p. 1091-1104, 2012.

Li X., Wang D., Li J., Zhang B., « Video search in concept subspace : A text-like paradigm », *In Proc. of CIVR*, 2007.

Platt J., ≪ Probabilistic outputs for support vector machines and comparison to regularize likelihood methods ≫, *in* A. Smola, P. Bartlett, B. Schoelkopf, D. Schuurmans (eds), *Advances in Large Margin Classifiers*, p. 61-74, 2000.

Safadi B., Derbas N., Hamadi A., Thollard F., Quénot G., Delhumeau J., Jégou H., Gehrig T., Kemal Ekenel H., Stifelhagen R., ≪ Quaero at TRECVID 2012 : Semantic Indexing ≫, *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, 2012.

Safadi B., Quénot G., ≪ Evaluations of multi-learner approaches for concept indexing in video documents ≫, *RIAO*, p. 88-91, 2010.

Salton G., Fox E. A., Wu H., ≪ Extended Boolean information retrieval ≫, *Commun. ACM*, vol. 26, n° 11, p. 1022-1036, November, 1983.

Smeaton A. F., ≪ Independence of contributing retrieval strategies in data fusion for effective information retrieval ≫, *Proceedings of the 20th Annual BCS-IRSG conference on Information Retrieval Research*, IRSG'98, British Computer Society, Swinton, UK, UK, p. 12-12, 1998.

Snoek C. G., Huurnink B., Hollink L., de Rijke M., Schreiber G., Worring M., ≪ Adding Semantics to Detectors for Video Retrieval ≫, *Trans. Multi.*, vol. 9, n° 5, p. 975-986, August, 2007.

Tiberius Strat S., Benot A., Bredin H., Qunot G., Lambert P., ≪ Hierarchical late fusion for concept detection in videos ≫, *ECCV 2012, Workshop on Information Fusion in Computer Vision for Concept Recognition*, Firenze, Italy, Oct., 2012.

Wang G., Forsyth D. A., ≪ Joint learning of visual attributes, object classes and visual saliency ≫, *ICCV'09*, p. 537-544, 2009.

Wei X.-Y., Jiang Y.-G., Ngo C.-W., ≪ Concept-Driven Multi-Modality Fusion for Video Search ≫, *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, n° 1, p. 62-73, 2011.

Weiss G., Provost F., The Effect of Class Distribution on Classifier Learning : An Empirical Study, Technical report, 2001.

Wu P., Dietterich T. G., ≪ Improving SVM Accuracy by Training on Auxiliary Data Sources ≫, *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, ACM, New York, NY, USA, p. 110, 2004.

Yan R., Hauptmann A. G., ≪ The combination limit in multimedia retrieval ≫, *In Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA ́03*, p. 339-342, 2003.

Zadeh L. A., ≪ Fuzzy sets ≫, *Information and Control*, vol. 8, p. 338-353, 1965.