Reclassement sémantique pour l'indexation de documents multimédia.

Abdelkader Hamadi [1]

1. UJF-Grenoble;1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France . Abdelkader.Hamadi@imag.fr

RÉSUMÉ. *Cet article décrit une nouvelle approche pour indexer des documents multimédia (vidéo avec son) par des concepts visuels. En plus des informations relatives au concept cible, l'idée développée propose d'intégrer la détection d'un ensemble d'autres concepts. L'avantage escompté par une telle combinaison est d'améliorer la performance d'un système d'indexation profitant des relations entre les concepts. Des expérimentations sur le corpus TRECVID 2012 sont présentées et commentées. Notre méthode a permis d'améliorer significativement les performances d'un bon système, jusqu'à +6% sur la précision moyenne.*

ABSTRACT. *This paper describes a new approach for indexing multimedia documents (video with sound) by visual concepts. In addition to information regarding a target concept, the idea developed proposes to integrate the detection of a set of other concepts. The expected benefit from such a combination is to consider the relationships between concepts in order to reclassify the results of an initial indexing system. Experiments on the TRECVID 2012 corpus are presented and discussed. Our method has significantly improved the performance of an already good system, up to +6 % on average precision.*

MOTS-CLÉS : *Indexation, Recherche, Multimédia, Détection de concepts visuels, Fusion.*

KEYWORDS: *Indexing, Retrieval, Multimedia, Visual concept detection, Fusion, TRECVID.*

1. Directeur de thèse : Georges Quénot. Co-directeur de thèse : Philippe Mulhem

1. Introduction

Nous nous intéressons dans ce travail à la détection automatique d'un grand nombre de concepts visuels. Le but est d'améliorer un système d'indexation existant, quelque soit sa performance. Nous proposons donc une approche basée sur un reclassement des vidéos/images, qui prend en compte deux types d'informations : des informations relatives au concept à détecter et celles concernant d'autres concepts.
Cet article se décompose de la manière suivante. Dans la section 2, nous décrivons certains travaux qui ont une relation avec notre proposition. Nous détaillons notre approche dans la section 3. Les expérimentations et les résultats sont présentés en section 4. Nous concluons dans la section 5.

2. État de l'art

L'annotation d'échantillons multimédia grâce à une connaissance liée à plusieurs concepts a déjà été étudiée par le passé (R Naphade *et al.,* 2002, Kennedy, 2007, Snoek *et al.,* 2006). Par exemple dans le cas de la recherche de vidéos, (Kennedy, 2007) propose d'utiliser les scores d'annotation de l'ensemble des images renvoyées pour la requête, et de les fusionner avec de nouveaux scores, pour former une nouvelle liste d'images réponses. Une autre classe de travaux s'intéresse à la prise en compte des relations entre concepts qui annotent les échantillons. (R Naphade *et al.,* 2002) proposent un multinet probabiliste bayésien pour modéliser explicitement les relations entre concepts via un graphe construit à base d'une ontologie. (Jiang *et al.,* 2007) utilisent des champs de Markov cachés pour calculer la probabilité finale pour qu'un concept annote une image, en tenant compte des probabilités d'occurrence d'autres concepts. (Bannour *et al.,* 2012) utilisent pour leur part, une hiérarchie de concepts pour affiner les annotations en fusionnant des classificateurs. Une approche qui est relativement similaire aux travaux présentés ici a été proposée dans (Snoek *et al.,* 2006), dans laquelle une approche (baptisée *charm*) consiste à créer un vecteur de 101 dimensions par plan vidéo, une dimension correspondant au score d'un concept pour le plan considéré. Ensuite, un classifieur est appris sur ces vecteurs et les annotations manuelles initiales. Les résultats obtenus n'ont pas été concluants, mais notre proposition se comporte beaucoup mieux, comme nous le montrons dans la suite.

3. Reclassement sémantique par regroupement

L'approche que nous proposons et que nous appelons dans la suite *reclassement sémantique par regroupement*, est basée sur le constat suivant : en raison de la richesse du contenu d'une image/vidéo en termes de sémantiques, tenter de détecter un concept visuel seul est une idée très naïve. En effet, les concepts n'existent pas isolément, certains concepts cooccurent toujours (*Animal* & *Vehicule*), certains autres, très souvent (*Sky* & *Airplane*). La présence de certains concepts exclut l'occurrence de certains autres (*Indoor* & *Outdoor*). Nous supposons qu'utiliser des informations

liées à d'autres concepts permettrait d'améliorer les performances initiales de détection d'un concept cible, par rapport au cas où aucune autre information externe n'est utilisée. Sur la base de cette idée, si l'on considère que nous avons un score de détection pour chaque concept, et en considérant pour chaque échantillons les scores de détection d'un ensemble de concepts comme vecteur caractéristique, nous pensons que les échantillons positifs se réuniront dans l'espace, et un échantillon négatif sera plus éloigné (qu'un exemple positif) des centres des groupes qui contiennent beaucoup d'exemples positifs. Cela conduit à notre proposition. Nous notons dans ce qui suit :

– $Score_{init}(e, c)$: le score de détection d'un concept c dans un échantillon e sans reclassement sémantique. On notera $Score_{init}(e, .)$ le vecteur contenant le score de détection de tous les concepts dans l' échantillon e.

– $Score_{recl}(e, c)$: le score de reclassement sémantique pour le concept c dans e ;

– $Score_{final}(e, c)$: le score final de détection, résultat d'application d'une fonction G de $R \times R$ dans R : $Score_{final}(e, c) = G(Score_{init}(e, c), F_{retro}(e, c))$

Notre approche modifie les scores d'une première classification $Score_{init}(e, c)$ pour calculer de nouvelles valeurs d'annotation automatique ($Score_{recl}(e, c)$). L'annotation d'un échantillon par un concept est décrite par les étapes suivantes :

1) Calculer de $Score_{init}(e, .)$ pour tous les échantillons ;

2) Faire un regroupement (*clustering*) suivant $Score_{init}(e, .)$ sur l'ensemble des échantillons d'apprentissage, annotés positivement ou négativement par c. Le résultat est un ensemble de CL_c clusters notés $clus_{c,j}$ avec $j \in [1, CL_c]$. Dans notre approche, un regroupement à base de K-moyennes (*K-means*) est utilisé, et le nombre CL_c est un paramètre optimisé sur un corpus de développement ;

3) Estimer la probabilité d'occurence du concept c dans chaque cluster $clus_{c,j}$:

$$P_c(+|clust_{c,j}) = \frac{\#\ positifs\ pour\ c\ dans\ clus_{c,j}}{\#\ positifs\ ou\ négatifs\ pour\ c\ dans\ clus_{c,j}} \quad [1]$$

4) Calculer la distance séparant $Score_{init}(e, .)$ des centroïdes des clusters $clust_{c,j}$, notée $dist(e, clus_{c,j})$. Notre choix s'est porté sur la distance L_1 (dite de Manhattan) normalisée par le nombre de dimensions considéré, dans notre cas $|C|$;

5) En déduire la valeur $Score_{recl}(e, c)$. Nous avons choisi d'utiliser un calcul de plus proches voisins, avec un paramètre K qui dénote le nombre des plus proches clusters considéré et N un facteur de normalisation :

$$Score_{recl}(e, c) = \frac{1}{N} * \sum_{j \in k\ plus\ proches\ voisins} \frac{P_c(+|clust_{c,j})}{dist(e, clust_{c,j})} \quad [2]$$

$$\text{avec } N = \sum_{j=1}^{CL_c} \frac{P_c(+|clust_{c,j})}{dist(e, clust_{c,j})}$$

6) Fusionner les scores de reclassement et les scores initiaux. Après avoir testé plusieurs fonctions G, nous avons opté pour une combinaison linéaire pondérée :

$Score_{final}(e,c) = \alpha_c.Score_{init}(e,c) + (1-\alpha_c).Score_{recl}(e,c)$
où α_c est un facteur de pondération

4. Expérimentations et résultats

4.1. *Données*

Notre évaluation a été réalisée sur les collections de données TRECVID 2012. Nous avons considéré un lexique de 346 concepts. Nos expérimentations ont été faites sur trois corpus : apprentissage, validation et test. Les annotations ont été fournies par l'annotation collaborative TRECVID 2012 (Ayache *et al.,* 2008). Nous avons utilisé la précision moyenne (MAP) comme mesure de performance, calculée sur les 346 concepts pour le corpus de validation, et sur 46 concepts pour le corpus de test. Ce choix est dû au fait que c'est la procédure *officielle* d'évaluation de TRECVID, et aussi parce que nous ne disposions pas des annotations des autres concepts.

4.2. *Expérimentations*

20 types de descripteurs sont utilisés. Ces descripteurs portent sur les textures/couleurs, SIFT, STIP, VLAD, VLAT, Percepts, ... 100 variantes au total de ces descripteurs (par exemple en fonction de la taille du dictionnaire pour les "bags of word") sont considérées. L'ensemble des descripteurs utilisé est décrit dans (Ballas *et al.,* 2012). Nous pouvons diviser nos expérimentations en trois étapes :

a) Détection initiale des concepts
En raison de leurs bons résultats, nous avons choisi d'utiliser MSVM (Safadi *et al.,* 2010) et KNN (Yang *et al.,* 2008) comme classifieurs supervisés initiaux. Comme entrée de ces détecteurs, les descripteurs décrits ci-dessus ont été utilisés. Pour chaque paire (concept, descripteur), les expérimentations suivantes ont été effectuées :1) Optimiser les paramètres sur l'ensemble d'apprentissage ; 2) Application sur le corpus de validation et évaluation sur 346 concepts ; 3) Application sur le corpus de test et évaluation sur 46 concepts (procédure officielle de TRECVID2012).

b) Fusion
Une fusion tardive des scores obtenus dans la première étape est réalisée afin d'améliorer les performances. Nous avons considéré trois résultats de fusion comme systèmes de base : 1) $Fusion\_1$: fusion basée sur un apprentissage 2) ; $Fusion\_2$: une fusion hiérarchique, décrite dans (Safadi *et al.,* 2012) ; 3) $Fusion\_3$: résultat de la fusion de $Fusion\_1$ et $Fusion\_2$.

c) Reclassement par regroupement
$Fusion\_1$, $Fusion\_2$ et $Fusion\_3$ donnent de bonnes valeurs de MAP, qui s'élèvent à plus de 0.2, ce qui donnerait un bon classement officiel à TRECVID. Dans nos expérimentations, une optimisation globale des paramètres (des classifieurs, CL_c, α_c) a été tentée, ce qui a donné des résultats décevants en terme de MAP. Cela s'explique,

	Corpus de validation		Corpus de Test	
	MAP initiale	MAP (gain %) après reclassement	MAP initiale	MAP (gain %) après reclassement
$Fusion\_1$	0.2010	0.2139 (+6.42)	0.2431	0.2522 (+3.75)
$Fusion\_2$	0.2469	0.2525 (+2.27)	0.2600	0.2591 (-0.34)
$Fusion\_3$	0.2488	0.2538 (+2.01)	0.2749	0.2774 (+0.90)

Tableau 1. *Résultats sur la collection TRECVID. Les paramètres sont optimisés sur le corpus d'apprentissage.*

d'une part, par la différence du nombre et des instances des exemples positifs et négatifs pour les différents concepts, et d'autre part, par la différence des performances obtenues pour les différents descripteurs. En effet, ces différences mènent sans doute à un regroupement différent avec les k-means. On a donc dans un second temps opté pour une optimisation locale pour chaque paire (concept, descripteur), choix pour lequel les résultats qui suivent sont présentés.

4.3. *Résultats*

Le tableau 1 montre les résultats obtenus en utilisant le reclassement sémantique par regroupement, sur les deux corpus de validation et de test. On remarque que notre approche améliore les résultats sur le corpus de validation, quelque soient les résultats initiaux utilisés. Le gain va entre +2.01% pour $Fusion_3$ et +6.42% pour $Fusion_1$. Cette différence de gain peut être expliquée par la différence des performances de la première classification : il est plus difficile d'améliorer un bon système qu'un moins bon. On notera que cette amélioration est très significative, selon le test de Student bilatéral par paires, où les valeurs de p sont inférieures à 3,1E-14. Les remarques sont similaires pour la collection de test, à part une legère dégradation obtenue en utilisant les scores $Fusion_2$. L'amélioration n'est pas aussi importante que dans le cas du corpus de validation. Ceci s'explique par le fait qu'il n'y a que 46 concepts dans la procédure officielle d'évaluation de TRECVID.

5. Conclusion

Nous avons proposé dans cet article une approche de reclassement sémantique pour détecter automatiquement des concepts dans des échantillons multimédia. Notre méthode repose sur l'utilisation de la détection d'un grand nombre de concepts, en modifiant les scores d'une première classification, en se basant sur le résultat d'un clustering d'information sémantique. Nous avons expérimenté notre proposition sur la collection de documents vidéos TRECVID 2012, en utilisant des caractéristiques provenant du consortium IRIM et de QUAERO. Les résultats obtenus montrent que

notre proposition améliore la qualité d'annotation des plans vidéos, de manière significative sur l'ensemble de développement de 346 concepts. A l'avenir, nous voulons étendre notre proposition à la prise en compte non pas des scores d'annotation automatique de tous les concepts à la fois, mais uniquement de certains concepts regroupés en familles. Cette approche devrait encore améliorer les résultats obtenus.

6. Remerciement

Ce travail a été partiellement réalisé dans le cadre du programme Quaero qui est financé par OSEO, l'organisme d'état français pour l'innovation.

7. Bibliographie

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., Benoît 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, nov, 2012.

Bannour H., Hudelot C., « Hierarchical image annotation using semantic hierarchies », *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, ACM, New York, NY, USA, p. 2431-2434, 2012.

Jiang W., Chang S.-F., Loui A., « Context- Based Concept Fusion with Boosted Conditional Random Fields », *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, p. I-949 -I-952, april, 2007.

Kennedy L. S., « A Reranking Approach for Context-based Concept Fusion in Video Indexing and Retrieval », *In Conference on Image and Video Retrieval*, 2007.

R Naphade M., Kozintsev I. V., Huang T. S., « Factor graph framework for semantic video indexing », *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 12, n° 1, p. 40-52, January, 2002.

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, nov, 2012.

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

Snoek C. G. M., Gemert J. C. V., Gevers T., Huurnink B., Koelma D. C., Liempt M. V., Rooij O. . D., Seinstra F. J., Smeulders A. W. M., Thean A. H. C., Veenman C. J., Worring M., « The MediaMill TRECVID 2006 semantic video search engine », *In Proceedings of the 4th TRECVID Workshop*, 2006.

Yang J., Hauptmann A. G., « (Un)Reliability of video concept detection. », *CIVR'08*, p. 85-94, 2008.