HAL
open science

Apprentissage actif avec une méthode de réordonnancement pour l'indexation et la recherche de vidéos

Bahjat Safadi, Georges Quénot

To cite this version:

Bahjat Safadi, Georges Quénot. Apprentissage actif avec une méthode de réordonnancement pour l'indexation et la recherche de vidéos. CORIA 2011 - COnférence en Recherche d'Information et Applications, Mar 2011, Avignon, France. pp.231-245. hal-00764524

HAL Id: hal-00764524

https://hal.science/hal-00764524

Submitted on 13 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Apprentissage Actif avec une Méthode de Réordonnancement pour l'Indexation et la Recherche de Vidéos

Bahjat Safadi et Georges Quénot

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

Bahjat.Safadi@imag.fr

RÉSUMÉ. *La recherche de vidéos peut être faite en ordonnant les échantillons en fonction de scores de probabilité produits par des classifieurs. Il est souvent possible d'améliorer la performance des systèmes par un réordonnancement de ces échantillons. Dans cet article, nous proposons une telle méthode et nous proposons également la combinaison de cette méthode avec un apprentissage actif pour l'indexation de vidéos. Les résultats expérimentaux montrent que la méthode de réordonnancement proposée a été en mesure d'améliorer la performance du système avec une augmentation d'environ 16-22% du score en moyenne sur la tâche d'indexation sémantique en TRECVID 2010. En outre, elle a amélioré la performance du système d'indexation des vidéos par apprentissage actif, en considérant l'aire sous la courbe (AUC) comme mesure d'évaluation de la performance de l'apprentissage actif. Notre méthode de réordonnancement améliore la performance d'environ 20% en moyenne sur la collection TRECVID 2007.*

ABSTRACT. *Video retrieval can be done by ranking the samples according to their probability scores that were produced by classifiers. It is often possible to improve the retrieval performance by re-ranking the samples. In this paper, we proposed such a method and we combined this method with active learning for video indexing. Experimental results showed that the proposed re-ranking method was able to improve the system performance with about 16-22% in average on TRECVID 2010 semantic indexing task. Furthermore, it improved significantly the performance of the video indexing system based active learning; by considering the Area Under Curve (AUC) as a metric measure for the performance of the active learning, our re-ranking method improved the performance with about 20% in average on TRECVID 2007.*

MOTS-CLÉS : *Indexation et de recherche des vidéos, Apprentissage Actif, Réordonnancement*

KEYWORDS: *Video Indexing and Retrieval, Active Learning, Re-ranking*

1. Introduction

L'indexation sémantique a constitué un champ de recherche très actif des dernières années. Elle consiste à construire automatiquement une description des vidéos pour en permettre une recherche par le contenu. Malgré le nombre important de travaux sur le sujet, l'indexation par le contenu de larges collections de vidéos autorisant la recherche de vidéos ou la navigation au sein d'une collection est toujours un problème ouvert.

L'indexation sémantique est généralement effectuée par apprentissage supervisé : un système de classification est entraîné à partir d'exemples positifs et négatifs d'un concept cible (dans un ensemble de développement). Un modèle est construit et est ensuite utilisé pour prédire la présence du concept donné dans de nouveaux échantillons (dans un ensemble de test). La prédiction est faite sous la forme d'un score homogène à la probabilité pour chaque échantillon de contenir le concept cible. Ces scores sont utilisés pour ordonner les échantillons de l'ensemble de test. Ils sont calculés indépendamment pour chaque échantillon sur la base des informations disponibles dans l'ensemble de développement. Il est souvent possible d'améliorer l'indexation ou le recherche en réordonnant les échantillons de l'ensemble de test en considérant le classement initial et la structure de cet ensemble.

En pratique, le volume de données qui peut être annoté manuellement pour l'entraînement des systèmes est limité. Une stratégie consiste à sélectionner les échantillons à annoter pour que l'annotation soit faite sur des données les plus utiles possibles (Angluin, 1988). L'apprentissage actif peut être utilisé pour choisir ces échantillons à annoter. Un utilisateur étiquette subjectivement un échantillon comme positif ou négatif. L'ensemble de ces échantillons est alors utilisé pour entraîner un classifieur qui sera ensuite utilisé pour faire une prédiction pour les échantillons non encore annotés. Les échantillons ayant les scores les plus élevés ou les plus incertains sont sélectionnés comme les plus informatifs pour le système et proposés à l'annotation.

Très souvent, les données utilisées pour l'indexation de vidéos sont très déséquilibrées, avec une proportion très faible d'échantillons positifs pour un concept donné. Cette caractéristique est particulièrement problématique dans le cadre de l'apprentissage supervisé classique. Dans le contexte des campagnes d'évaluation TRECVID par exemple, la classe minoritaire représente typiquement moins de 1% (Smeaton *et al.,* 2006). Une approche proposée pour gérer le problème des données non équilibrées consiste à générer un ensemble équilibré en sous-échantillonnant la classe majoritaire dans l'ensemble de développement (Bishop, 2007). De plus, il est possible de compenser la perte d'information induite par le sous-échantillonnage de la classe majoritaire par la génération de plusieurs ensemble de données équilibrés, et de fusionner les sorties produites par les classifieurs construits sur ces ensembles. Une combinaison de l'apprentissage actif avec cette approche à classifieurs multiples permet de significativement améliorer l'efficacité de l'annotation multimédia, en particulier dans le cadre de données fortement déséquilibrées comme celle disponibles dans TRECVID (Safadi *et al.,* 2010).

Les performances d'un système d'indexation vidéo basé sur l'apprentissage actif peuvent aussi être améliorées en réordonnant les résultats obtenus après chaque itération d'apprentissage actif. Récemment, plusieurs méthodes de réordonnancement ont été proposées. Nous en présentons certaines maintenant.

Context fusion (Wei Jiang et al., *2007, Liu* et al., *2007)* : les résultats de différents type de requête (requête à base de concept, requête textuelle, requête par exemple) sont utilisés pour réordonner la liste des résultats, en se concentrant sur la fusion des sorties des différents modèles. Cette méthode nécessite d'apprendre de nouveaux classifieurs et de nouveaux descripteurs. Comme nous utilisons aussi la fusion des sorties obtenues par différents modèles, nous considérons cette approche comme notre modèle de référence.

Classification-based re-ranking (Kennedy et al., *2007)* : les résultats obtenus par des systèmes basiques sont utilisés pour découvrir les cooccurrences de motifs entre les cibles sémantiques et les caractéristiques extraites. Cette approche est similaire à l'approche *"learning to rank"* (Herbrich *et al.,* 1999), dans laquelle on apprend un modèle qui permet de prédire directement une liste triée de résultats. Dans (Kennedy *et al.,* 2007), les auteurs utilisent les exemples en début de liste (respectivement fin de liste) comme exemples pseudo positifs (respectivement pseudo négatifs) pour entraîner un nouveau modèle. Lorsque le classifieur est un SVM, la méthode est appelée *RankSVM* (Herbrich *et al.,* 1999).

Ordinal re-ranking (Yang et al., *2008)* : les auteurs réordonnent les résultats initiaux en utilisant des cooccurrences de motifs, ce qui produit des scores de réordonnancement. Les scores finaux sont une combinaison pondérée des scores initiaux et des scores de réordonnancement. Les auteurs adoptent une méthode d'apprentissage pour réordonner certains concepts ; l'algorithme de réordonnancement est utilisé pour réordonner le reste des concepts.

Dans le cadre de collections vidéo, l'unité considérée n'est pas en général la vidéo complète car elle présente une granularité qui n'est pas intéressante pour les besoins des utilisateurs finaux. L'unité d'indexation et de recherche est souvent en pratique le plan vidéo.

La contribution présentée dans ce papier est une méthode de réordonnancement d'une liste des plans vidéo en utilisant les scores obtenus par un classifieur initial et la connaissance de la vidéo et de sa nature. Notre travail est proche de (Wang *et al.,* 2009) où les auteurs réordonnent les résultats des recherches précédentes en utilisant la moyenne des scores des plans de la même vidéo. Nous approfondissons cette approche en considérant une moyenne généralisée des scores des plans vidéo. Appliquée dans le cadre de l'apprentissage actif pour l'indexation de vidéo, cette méthode permet d'obtenir des gains significatifs.

Le papier est organisé comme suit : nous présentons notre méthode de réordonnancement en section 2 ; l'apprentissage actif avec réordonnancement est présenté en

section 3. La section 4 présente les résultats expérimentaux. Nous concluons enfin en section 5.

2. Méthode de réordonnancement

Un système de recherche d'information multimédia doit trier les plans vidéo selon une estimation de leur pertinence, pertinence relative à ce que l'utilisateur a envie de voir. Cette estimation peut être basée sur les scores de prédiction des classifieurs, comme une probabilité pour un plan de contenir le concept cible. En règle générale, les listes triées produites par les différents classifieurs contiennent des erreurs. Une méthode de réordonnancement permettra alors de minimiser l'erreur présente dans les listes triées.

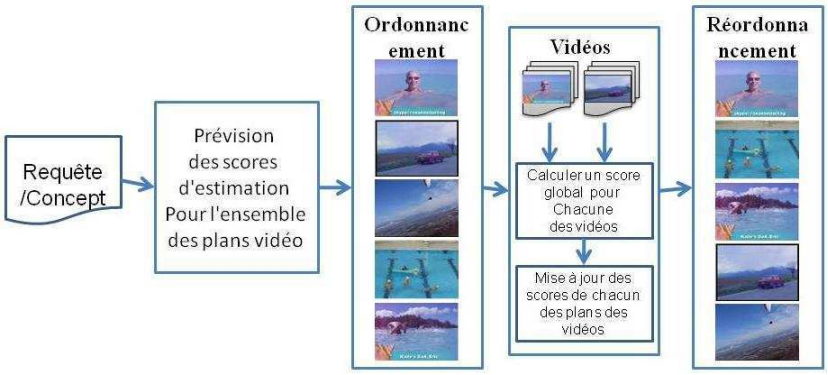

Figure 1. *Le cadre général de notre système de réordonnancement.*

La méthode que nous proposons part de l'hypothèse selon laquelle les vidéos ont un contenu relativement homogène. En conséquence, la présence d'un concept donné dépend beaucoup de la nature de la vidéo elle même. Cependant, les scores, homogènes à des probabilités de présence d'un concept, sont estimés indépendamment pour chaque plan. La méthode de réordonnancement des plans vidéos s'effectue en deux étapes :

1) pour chaque vidéo, un score global de présence du concept dans la vidéo est calculé. Ce score est calculé en prenant en compte les scores de tous les plans de la vidéo ;

2) le score de chaque plan est réévalué en prenant en compte le score global de la vidéo auquel il appartient.

Considérons la collection de test qui consiste en un ensemble de vidéos $V = (v_1, v_2, \ldots, v_m)$, m étant le nombre de vidéos dans la collection. Chaque vidéo v_i est composée d'une séquence de plans $v_i = [s_{i1}, s_{i2}, \ldots, s_{in_i}]$, avec n_i le nombre de plans de la vidéo v_i. Pour chaque plan, s_{ij}, on dispose d'un score de classification initial x_{ij}, qui est calculé, dans notre cas, par apprentissage supervisé sur l'ensemble

de développement. Le cadre général de l'algorithme de réordonnancement est proposé en figure 1.

Plusieurs options - parmi lesquelles *les moyennes arithmétique, géométriques ou harmonique, le minimum, le maximum, ...* - sont possibles pour le calcul des scores globaux z_i pour la vidéo v_i. Toutes ces moyennes peuvent être obtenues à partir de la moyenne généralisée (équation 1) en modifiant les valeurs du paramètre α :

$$z_i = \left(\frac{\sum_{j=1}^{n_i} (x_{ij})^\alpha}{n_i} \right)^{1/\alpha} \qquad [1]$$

$\alpha \to -\infty$: minimum ;
$\alpha = -1$: moyenne harmonique ;
$\alpha \to 0$: moyenne géométrique ;
$\alpha = +1$: moyenne arithmétique ;
$\alpha = +2$: racine carrée de la moyenne des carrés ;
$\alpha \to +\infty$: maximum.

Le paramètre α devra être estimé par validation croisée au sein de l'ensemble de développement.

Une fois le score global z_i calculé, le score de chaque plan est réévalué en fonction de son score précédent et du score global de la vidéo à laquelle il appartient. Ici encore, plusieurs options s'ouvrent à nous. Notre choix se porte sur une fusion multiplicative pondérée :

$$x'_{ij} = x_{ij}^{1-\gamma} \times z_i^\gamma, \qquad [2]$$

où γ est un paramètre qui contrôle la « force » du réordonnancement. Ce paramètre devra, lui aussi, être estimé par validation croisée au sein de l'ensemble de développement.

3. Apprentissage actif avec réordonnancement

L'apprentissage actif est une technique utilisée pour sélectionner *de manière optimale* des échantillons à annoter pour les ajouter à l'ensemble d'apprentissage. Il a été originalement modélisé et décrit par Cohn, Ghahramani, and Jordan dans (Cohn *et al.,* 1996), où ils appliquent le principe à des approches d'apprentissage statistique.

Beaucoup de problèmes pratiques peuvent bénéficier de l'apprentissage actif. L'apprentissage actif a été appliqué dans un système de recherche de plans vidéos (Ayache *et al.,* 2007) ; Dans un tel système, l'apprentissage actif est utilisé pour

sélectionner les plans qui devront être annotées manuellement. Les plans sont sélectionnés selon leurs capacité à augmenter la connaissance du système[1]. L'algorithme d'apprentissage actif demande dynamiquement à l'annotateur son jugement pour un exemple dûment sélectionné, c'est à dire de dire pour un concept donné et un plan vidéo donné si le concept est ou non visible dans le plan. Dans la mesure où le système choisit les exemples à annoter, le nombre d'exemples nécessaires pour l'apprentissage d'un concept donné est en général bien inférieur à celui classiquement requis dans le cadre d'un apprentissage supervisé classique. L'introduction de techniques de réordonnancement peut améliorer encore les performances de ces systèmes. Elles peuvent être utilisées pour réordonner les résultats à chaque itération de la procédure d'apprentissage actif, permettant de faire apparaitre en début de liste les éléments les plus informatifs pour le système.

Algorithm 1 Apprentissage actif avec réordonnancement

V l'ensemble des vidéos de la collection
S : ensemble des plans vidéos de la collection.
L_i, U_i : ensemble de plans étiquetés, vs non étiquetés.
A=(Train, Predict) : Algorithme de classification de base.
Q : fonction de sélection (ou de requête).
Initialise L_i (*e.g.* 10 positifs & 20 négatifs).
while $S \setminus L_i \neq \emptyset$ **do**
  $C \leftarrow$ Train(L_i)
  $X \leftarrow$ Predict(S,C) // produit les scores x_{ij}
  for all $v_i \in V$ **do**
    Calculer le score global z_i pour v_i (eq. 1)
    for all $s_{ij} \in v_i$ **do**
      Mise à jour des scores des plans s_{ij} (eq 2)
    end for
  end for
  Appliquer Q sur X et sélectionner $\tilde{x} \in U_i$ échantillons.
  $\tilde{y}$ = Annoter($\tilde{x}$)
  $L_{i+1} \leftarrow L_i \cup (\tilde{x}; \tilde{y})$
  $U_{i+1} \leftarrow U_i \setminus \tilde{x}$
end while

L'algorithme d'apprentissage actif avec réordonnancement est détaillé par l'algorithme 1. À chaque itération i, l'ensemble de développement S est divisé en deux parties : L_i et U_i les exemples étiquetés et non étiquetés respectivement. Un classifieur C est appliqué sur L_i avec les annotations associées. Il est ensuite utilisé en prédiction sur l'ensemble S −l'ensemble des échantillons étiquetés et non étiquetés− afin de produire des scores. Dans le cas « classifieurs multiples »les prédictions des

1. Un plan déjà présent dans le système n'apportera aucune information supplémentaire. Un plan représentant un *outlier* sera également à éviter.

classifieurs élémentaires sont fusionnées pour produire un unique score de prédiction par échantillon. Une fois ces scores disponibles, les scores globaux Z_i peuvent être calculés pour chaque vidéo. Le score de chaque plan peut alors être réévalué en combinant les scores globaux et les scores des plans : pour la vidéo v_i, le score du plan j de la vidéo v_i sera obtenu par combinaison de Z_i, et de x_{ij}. Dans la mesure où les prédictions sont faites sur tous les échantillons de l'ensemble S, les vidéos dont les plans sont étiquetés comme positifs auront probablement un score Z qui donnera plus de poids à leurs plans. La fonction de sélection (ou requête) est alors appliquée sur les scores produits sur les échantillons de U_i ce qui permet de générer une liste triée d'exemples à indexer. Les premiers éléments de cette liste vont permettre de constituer l'ensemble $\tilde{x}$ des plans sélectionnés pour l'indexation : l'ensemble $\tilde{x}$, auquel sont associées les étiquettes $\tilde{y}$, est alors fusionné à L_i pour produire l'ensemble L_{i+1}. De plus U_{i+1} est constitué de $U_i \setminus \tilde{x}$.

Pour des raisons d'implantations, l'algorithme d'apprentissage de base A (dans notre cas un SVM) et la fonction de sélection Q doivent être fixés. D'autre paramètres doivent également être fixés, comme par exemple le ratio entre nombre d'exemples positifs et nombre d'exemples négatifs, la fonction de fusion utilisées pour combiner les résultats des différents classifieurs élémentaires, ou encore la méthode utilisée pour choisir le nombre de nouveaux échantillons que l'on doit intégrer à chaque itération.

4. Expérimentations

Dans cette section, nous présentons deux expériences réalisées : la première expérimentation présente la validation croisée et l'évaluation de la méthode de réordonnancement selon le protocole proposé dans la tâche d'indexation sémantique de la campagne d'évaluation TRECVID 2010. La seconde expérimentation présente l'évaluation de la technique d'apprentissage actif avec réordonnancement pour l'indexation de vidéo sur la collection TRECVID 2007.

4.1. *Réordonnancement sur la tâche d'indexation sémantique de TRECVID 2010*

Cette expérimentation a été conduite dans le cadre de tâche d'indexation sémantique de TRECVID 2010. Cette tâche consiste à rechercher des concepts visuels dans les plans vidéos. La collection comprend 400 heures de vidéos du Web faisant partie de la collection Internet Archive. Cette collection est décomposée en une partie test et une partie développement de 200 heures chacune. L'ensemble de développement contient 119 685 plans vidéo partiellement étiquetés (que nous avons divisé en 59 800 plans pour l'apprentissage et 59 885 pour la validation). L'ensemble de test contient 146 788 plans dont les étiquettes ne sont pas connues. La tâche est de produire une liste ordonnée de 2000 plans de l'ensemble de test pour chacun des 130 concepts pour lesquels des annotations ont été fournis sur l'ensemble de développement. L'évaluation est faite en calculant la moyenne des précisions moyennes (MAP)

sur 30 concepts choisis par les organisateurs. Une précision moyenne inférée est en pratique utilisée (Aslam *et al.,* 2006) ; elle constitue une très bonne approximation de la précision moyenne et peut être calculée à partir d'un nombre sensiblement plus faible de jugements.

L'algorithme de réordonnancement a été évalué en utilisant quatre résultats globaux de classification différents, obtenus par différentes stratégies de combinaisons de résultats intermédiaires : *Fusion_MAP*, *Fusion_OPT*, *Fusion_GA_MAP* et *Fusion_GA_OPT*. Ces stratégies de fusion ont été appliquées à des vecteurs de scores obtenus en entraînant différents systèmes sur 46 descripteurs différents (audio et vidéo). Ces descripteurs ont été produits par les partenaires du projet IRIM du GDR-ISIS (Gorisse *et al.,* 2010) et par l'Institut de Technologie de Karlsruhe. Chacune de ces stratégies de fusion peut être considérée comme le résultat de référence (*baseline*). La différence entre ces quatre combinaisons est la suivante : pour les versions *_MAP*, une pondération suivant la performance (MAP) estimée est utilisée pour la dernière étape de la fusion hiérarchique alors que pour les versions *_OPT*, une optimisation des pondérations est effectuée directement par validation croisée ; pour les versions incluant *_GA*, des algorithmes génétiques ont été utilisés lors de certaines étapes de la fusion.

4.1.1. *Optimisation des paramètres*

Comme mentionné en section 2, notre système nécessite le réglage de deux paramètres : α et γ. Ce réglage a été réalisé sur l'ensemble de développement de TRECVID 2010 en utilisant quatre soumissions précédemment mentionnées (*Fusion_**). Modifier le paramètre α dans l'équation 1 engendrera différentes fonctions de calcul du score global z. Nous avons considéré ici 6 valeurs pour α. Nous avons inclus deux valeur critiques pour le paramètre γ : $\gamma = 0$ qui est équivalent à « pas de réordonnancement », et $\gamma = 1$ qui attribue à tous les plans d'une même vidéo, le score global z.

Le réglage des paramètres est présenté en figure 2, où la performance du système (mesurée par la MAP) est calculée pour les 130 concepts sur l'ensemble de validation en utilisant la méthode *Fusion_MAP*. Comme on peut le voir, la fonction *maximum* n'est pas meilleure que la méthode de référence dans la mesure où, pour chaque vidéo, il y a au moins un score dont la valeur est proche de 1. La fonction *minimum* produit des résultats mauvais car chaque vidéo contient au moins un plan dont le score est proche de 0. La moyenne géométrique produit également de mauvais résultats. La moyenne harmonique se comporte bien pour de faibles valeurs de γ, mais les performances sont mauvaises lorsque γ augmente. Les meilleures performances sont obtenues par la moyenne et la racine carrée de la moyenne des carrés, la meilleure performance étant obtenue pour $\alpha = 2$ et $\gamma = 0.4$, à savoir racine de la moyenne des carrés et une importance presque égale du score du plan et du score global.

Afin de vérifier la stabilité du paramètre γ, nous avons fixé la valeur du paramètre α à 2 pour les quatre soumissions du projet IRIM. La figure 3 présente les résultats obtenus en faisant varier γ en utilisant les quatre méthodes (soumissions) précédemment

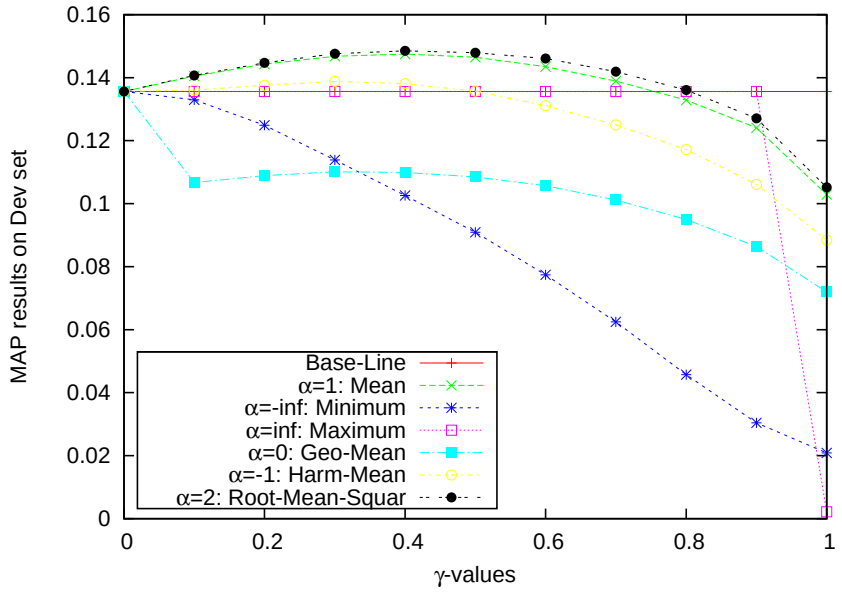

Figure 2. *Réglage des paramètres α et γ : réordonnancement des résultats après fusion sur une soumission avec différentes valeurs de α ; MAP sur les 130 concepts calculée sur l'ensemble de développement de TRECVID 2010.*

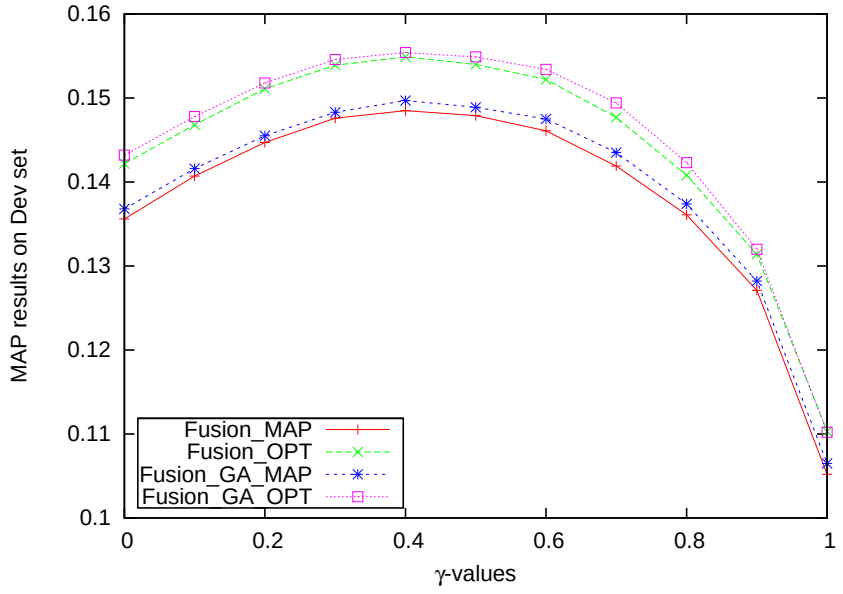

Figure 3. *Réglage du paramètre γ : réordonnancement des résultats après différentes stratégies de fusions ; MAP sur les 130 concepts calculée sur l'ensemble de développement de TRECVID 2010.*

mentionnées*Fusion_**. Comme on peut le constater sur la figure, le comportement de l'algorithme de réordonnancement est consistant entre les différentes soumissions. Ici encore, une valeur $\gamma = 0$ est équivalente à ne pas faire de réordonnancement. Comme précédemment, la meilleure valeur de γ est 0.4.

4.1.2. *Évaluation sur l'ensemble de test*

Une fois l'optimisation faite des paramètres, nous pouvons fixer leurs valeurs ($\alpha = 2$ et $\gamma = 0.4$) et faire les prédictions en aveugle sur l'ensemble de test. Nous comparons les résultats obtenus après réordonnancement à ceux obtenus initialement par les quatre méthodes de fusion. L'amélioration en terme de MAP est présentée dans le tableau 1 : sur cette collection, le gain obtenu par notre méthode va jusqu'à 22%. Ce gain sur l'ensemble de test est significativement plus important que celui que nous avons obtenu par validation croisée sur l'ensemble de développement. Cela tient peut-être au fait que les vidéos de l'ensemble de test sont en moyenne sensiblement plus courtes que celles de l'ensemble de développement et donc que l'hypothèse d'un contenu homogène y est donc souvent mieux satisfaite.

Pour aller plus loin dans l'analyse, le résultat de notre algorithme de réordonnancement est comparé à la méthode de base pour chacun des 30 concepts (figure 4). Tous les concepts ne sont pas affectés de la même manière par notre méthode. Il n'y a pas ou presque de cas où le réordonnancement conduit à une perte de performances mais le gain est très variable selon les concepts cibles. Pour environ 20% d'entre eux le gain est significatif et pour certains d'entre eux il est très important : *Animal* (143%) et *Swimming*(245%)). Conformément à notre hypothèse, ces concepts apparaissent bien groupés et dans un nombre relativement faible de vidéos.

Tableau 1. *Résultats obtenus sur l'ensemble de test ; MAP calculée sur 30 concepts*

Soumission	Fusion de base	réordonnancement	Gain%
Fusion_MAP	0.0476	0.0577	21
Fusion_GA_MAP	0.0479	0.0584	22
Fusion_OPT	0.0485	0.0563	16
Fusion_GA_OPT	0.0484	0.0568	17

4.2. *Évaluation de l'algorithme de réordonnancement combinée à l'apprentissage actif sur la collection TRECVID 2007*

La collection TRECVID 2010 ne pouvait pas être utilisée pour l'évaluation de la combinaison de l'apprentissage actif et de notre méthode de réordonnancement sur l'indexation de vidéos car il fallait une annotation complète de l'ensemble de développement pour pouvoir simuler le processus d'apprentissage actif et la collection de développement de TRECVID 2010 n'était annotée qu'à 35% en moyenne. Nous avons donc utilisé pour cette évaluation la collection TRECVID 2007 sur laquelle les

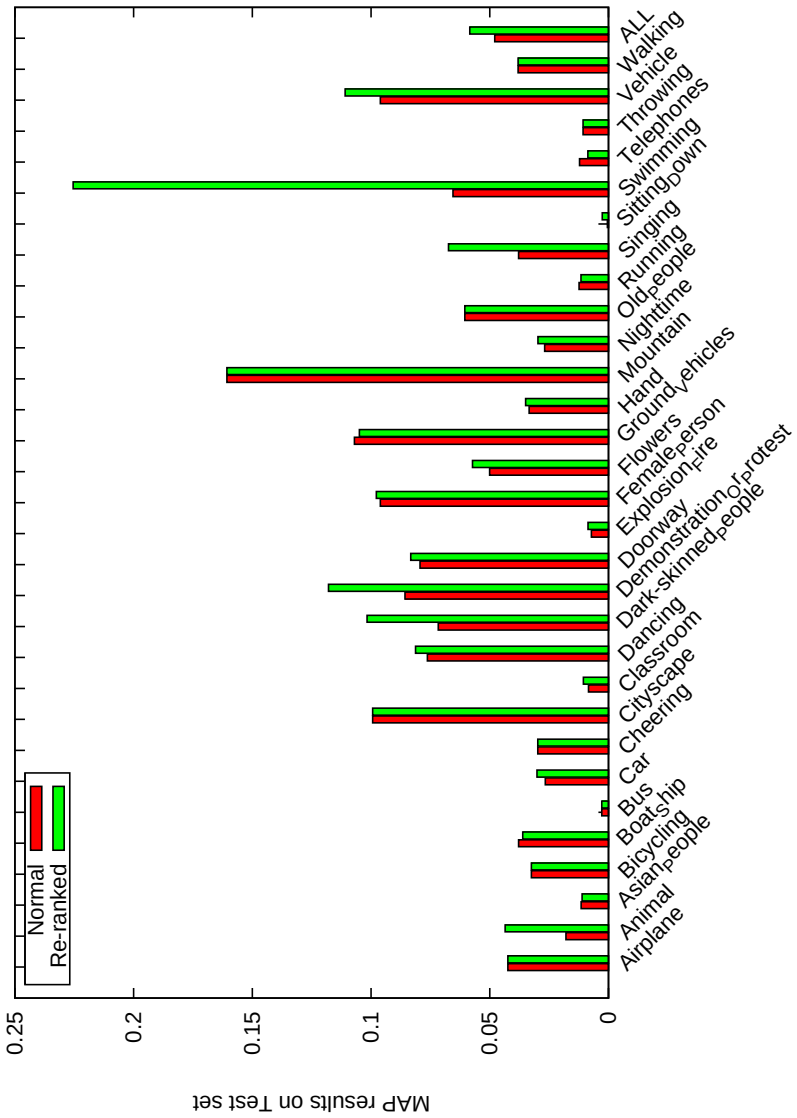

Figure 4. *Résultats de l'algorithme de réordonnancement sur 30 concepts de l'ensemble de test de TRECVID 2010.*

20 concepts de TRECVID 2008 ont été annotés en totalité. La collection TRECVID 2007 comprend 100 heures de films documentaires en Hollandais et elle se décompose en deux sous-ensembles de 50 heures chacun, pour le développement et pour le test, pour un total de 21 532 et 22 084 plans respectivement.

L'ensemble d'apprentissage étant complètement annoté, nous pouvons simuler l'apprentissage actif en incorporant au fur et à mesure les nouvelles annotations dans l'ensemble d'apprentissage. Nous pouvons donc observer l'évolution des performances obtenues sur l'ensemble de test en utilisant pour l'apprentissage des sous-ensembles croissants de l'ensemble de développement (sous-ensembles déterminés par l'apprentissage actif).

Quatre descripteurs d'images ont été utilisés. Ces descripteurs ont été produits par différents partenaires de l'action IRIM du GDR ISIS (Quénot *et al.,* 2009) :

– **CEALIST_global_tlep**, texture locale et histogramme RGB, 576 dimensions.
– **LEAR_bow_sift_1000**, histogramme de SIFT, 1000 dimensions.
– **ETIS_global_qwm1x3**, coefficients sur trois bandes, 96 dimensions.
– **LIG_hg104**, histogramme RGB et transformée de Gabor, 104 dimensions.

Nous avons utilisé les classifieurs à SVM multiple avec noyau RBF comme dans le cadre de (Safadi *et al.,* 2010). Nous avons conservé les valeurs $\alpha = 2$ et $\gamma = 0.4$ utilisées précédemment. Nous n'avons pas cherché de valeurs plus optimales pour cette collection et cette tâche car cela aurait trop lourd dans le cadre de l'apprentissage actif et d'après les expérimentations précédentes, le système n'est pas trop sensible à la valeur de ces paramètres. Le classifieur global utilisé pour l'apprentissage actif est une fusion tardive des classifieurs optimisés pour chacun des quatre descripteurs individuels.

4.2.1. *L'apprentissage actif et le démarrage à froid*

Lors du calcul du nombre d'exemples à annoter à chaque itération de l'apprentissage actif, nous avons utilisé un incrément variable. En pratique, nous avons utilisé 40 étapes au total, en considérant l'échelle géométrique suivante :

$$S_k = S_0 \times \left(\frac{N}{S_0}\right)^{k/K}$$

où N est la taille de l'ensemble de développement, S_0 est la taille de l'ensemble d'apprentissage au départ, K est le nombre total d'étapes, et k est l'étape courante. À chaque étape (ou itération), l'algorithme calcule S_k comme étant la taille du nouvel ensemble d'apprentissage, et choisit $S_k - S_{k-1}$ échantillons à étiqueter.

Dans cette évaluation, la moyenne harmonique a été utilisée comme stratégie de fusion intermédiaire, et nous avons utilisé l'échantillonnage pertinence en tant que stratégie pour l'algorithme d'apprentissage actif pour sélectionner de nouveaux échantillons d'être annoté. Le problème de l'ensemble de départ n'a pas été investigué et nous avons tiré au hasard 10 exemples positifs et 20 exemples négatifs.

4.2.2. *Performance de l'apprentissage actif*

La figure 5 présente les performances de l'apprentissage actif couplé à notre méthode de réordonnancement. La performance est mesurée par la MAP en fonction du nombre d'itérations de l'algorithme (40 au total). Etant donnée cette mesure, plus la pente de la courbe est importante, et plus le plateau est élevé, meilleure est la méthode. La courbe FUSION est la méthode de référence (pas de réordonnancement). Les deux autres courbes sont obtenues par réordonnancement. Elles diffèrent par l'ensemble sur lequel le réordonnancement est effectué : soit sur l'ensemble de test seulement (donc a posteriori et en dehors de l'apprentissage actif), soit à la fois sur l'ensemble et développement lors de l'apprentissage actif et aussi sur l'ensemble de test lors de la reconnaissance. Comme on peut le lire sur la courbe, le réordonnancement permet non seulement d'obtenir des performances similaires à celles obtenues en utilisant toutes les annotations en utilisant uniquement 3500 exemples annotés (soit 16% de la collection), mais il permet également d'obtenir des résultats encore meilleurs en utilisant environ 7000 exemples annotés (soit 32% de la collection). Le réordonnancement donne des résultats légèrement meilleurs lorsqu'il est appliqué à la fois sur l'ensemble de développement et sur l'ensemble de test.

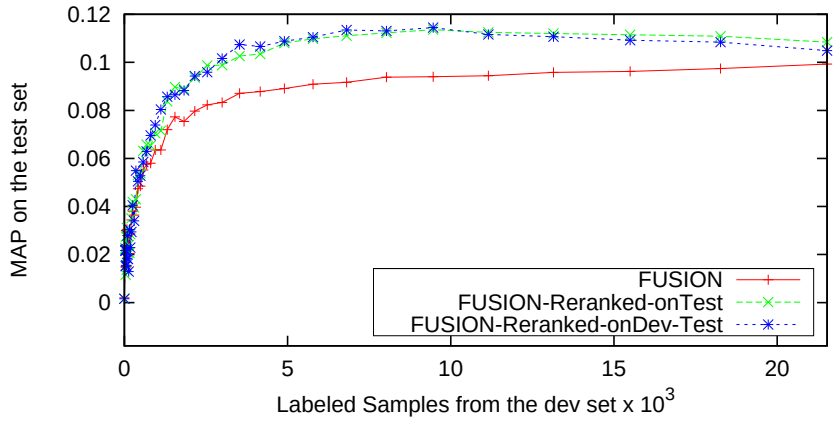

Figure 5. *Résultats obtenus par réordonnancement dans le cadre de l'annotation pour l'indexation de vidéos (collection test de TRECVID 2007).*

Nous voulons maintenant définir une mesure permettant de comparer deux méthodes d'apprentissage actif. Si A_i est la surface sous la courbe (AUC) i , alors $G_{a-b} = (A_a - A_b)/A_b$ permet de comparer deux stratégies d'apprentissage actif. Nous proposons de renormaliser les données en fonction de n le nombre total d'itérations, et (x_i, y_i) le nombre d'exemples annotés avec $(x_{n+1}, y_{n+1}) = (x_0, y_0)$:

$$A = \frac{1}{2}\left|\sum_{i=0}^{n+1} x_i \times y_{i+1} - y_i \times x_{i+1}\right|$$

Tableau 2. *Gain de performance obtenu par l'algorithme de réordonnancement.*

	(*AUC*)	Gain(%)
Sans réordonnancement	0.38	
Réordonnancement test	0.45	20
Réordonnancement dev et test	0.46	21

Le tableau 2 présente les gains obtenus par notre méthode de réordonnancement dans le contexte de l'apprentissage actif. Sur cette collection, avec les descripteurs utilisés, le gain est de l'ordre de 20%.

5. Conclusions

La recherche d'information dans des vidéos peut être réalisée via le classement des plans selon les scores de pertinences obtenus par des classifieurs. Il est souvent possible d'améliorer les performances des systèmes en réordonnant les plans *a posteriori*. Dans ce papier, nous avons proposé une méthode de réordonnancement qui améliore les performances de l'indexation sémantique de vidéos. De plus, cette technique de réordonnancement peut être combinée avec des techniques d'apprentissage actif. Les résultats expérimentaux proposés montrent que notre algorithme améliore significativement (de 16% à 20% en moyenne) les résultats de notre système lorsque appliqué à la tâche d'indexation sémantique de TRECVID 2010. Par ailleurs, le gain obtenu dans le cadre d'un système d'annotation basée sur l'apprentissage actif et du même ordre de grandeur : en considérant les surfaces sous les courbes comme mesure de performance, notre algorithme permet un gain de l'ordre de 20% sur TRECVID 2007.

Ce travail peut être poursuivi dans d'autres directions. En particulier, le réordonnancement se fait ici sur la base d'un score global sur l'ensemble de la vidéo. Ceci est bien adapté pour des vidéos courtes comme celles de TRECVID 2010 mais l'est peut-être moins dans le cas de vidéos plus longues et aussi selon leur type de contenu. Une possibilité serait de recalculer les scores des plans non pas en fonction de la vidéo complète mais en fonction d'un voisinage plus restreint dont la taille serait à optimiser.

Remerciements

Ce travail a été soutenu par le programme Quaero et par le projet IRIM du GDR ISIS. Les auteurs tiennent à remercier Franck Thollard pour des discussions sur des versions précédentes de ce papier.

6. Bibliographie

Angluin D., « Queries and Concept Learning », *Machine Learning*, vol. 2, p. 319-342, 1988.

Aslam J., Pavlu V., Yilma E., « Statistical Method for System Evaluation Using Incomplete Judgments », *Proceedings of the 29th ACM SIGIR Conference*, 2006.

Ayache S., Quénot G., « Evaluation of active learning strategies for video indexing », *Image Commun.*, vol. 22, n° 7-8, p. 692-704, 2007.

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

Cohn D. A., Ghahramani Z., Jordan M. I., « Active Learning with Statistical Models », *CoRR*, 1996.

Gorisse D., Precioso F., Gosselin P., Granjon L., Pellerin D., Rombaut M., Bredin H., Koenig L., Lachambre H., Khoury E. E., Vieux R., Mansencal B., Zhou Y., Benois-Pineau J., Jégou H., Ayache S., Safadi B., Tong Y., Thollard F., Quénot G., Benoît A., Lambert P., « IRIM at TRECVID 2010 : Semantic Indexing and Instance Search », *Proccedings of the workshop on TREC Video Retrieval Evaluation*, 15-17 Nov, 2010.

Herbrich R., Graepel T., Obermayer K., « Support Vector Learning for Ordinal Regression », *Ninth Intl. Conf. on Artificial Neural Networks*, p. 97-102, 1999.

Kennedy L. S., Chang S.-F., « A reranking approach for context-based concept fusion in video indexing and retrieval », *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, ACM, New York, NY, USA, p. 333-340, 2007.

Liu J., Lai W., Hua X.-S., Huang Y., Li S., « Video search re-ranking via multi-graph propagation », *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, ACM, New York, NY, USA, p. 208-217, 2007.

Quénot G., Delezoide B., le Borgne H., Moellic P., Gorisse D., Precioso F., Wang F., Merialdo B., Gosselin P., Granjon L., Pellerin D., Rombaut M., Bredin H., Koenig L., Lachambre H., Khoury E. E., Mansencal B., Benois-Pineau J., Jégou H., Ayache S., Safadi B., Fabrizio J., Cord M., Glotin H., Zhao Z., Dumont E., Augereau B., « IRIM at TRECVID 2009 : High Level Feature Extraction », *Proccedings of the workshop on TREC Video Retrieval Evaluation*, 16-17 Nov, 2009.

Safadi B., Quénot G., « Active learning with multiple classifiers for multimedia indexing », *Proceedings of the 8th workshop on Content Based Multimedia Indexing*, Grenoble, France, June, 2010.

Smeaton A. F., Over P., Kraaij W., « Evaluation campaigns and TRECVid », *MIR'06 : Proceedings of the 8th ACM International Worksh op on Multimedia Information Retrieval*, ACM Press, New York, NY, USA, p. 321-330, 2006.

Wang F., Merialdo B., « Eurecom at TRECVID 2009 High-Level Feature Extraction », *TREC2009 notebook*, 16-17 Nov, 2009.

Wei Jiang S.-F. C., Loui A. C., « Context-Based Concept Fusion with Boosted Conditional Random Fields », *ICASSP (1)*, p. 949-952, 2007.

Yang Y.-H., Hsu W. H., « Video search reranking via online ordinal reranking », *ICME*, p. 285-288, 2008.