Classification et Sélection de Caractéristiques basées sur les Concepts Sémantiques pour la Recherche d'Information Multimédia

Thierry Urruty*, Ismael Elsayad *, Adel Lablack*
Yue Feng**, Jose M. Joemon**

*University of Lille 1
59655 Villeneuve d'Ascq,France
{thierry.urruty, ismael.elsayad, adel.lablack}@lifl.fr
**Multimedia Information Retrieval Group
University of Glasgow, UK
{jj,yuefeng}@dcs.gla.ac.uk

Résumé. Le besoin récent de nombreuses applications multimédia basées sur le contenu a engendré une demande croissante de technologies dans le domaine de la recherche d'information multimédia. Basée sur l'état de l'art des techniques existantes, nous proposons dans cet article une approche de recherche d'information multimédia qui prend en compte les informations de scène et exploite un modèle de sélection de caractéristiques. Les principaux avantages de notre modèle de recherche par rapport aux modèles existants sont : (i) une méthode de classification basée sur des catégories de concept sémantique; (ii) un modèle de sélection de caractéristiques; (iii) un index multidimensionnel. Notre framework propose un bon compromis entre précision et rapidité de la recherche.

1 Introduction

De nos jours, la production de données multimédia est en forte croissance. Le domaine de la recherche d'information reçoit donc une attention très particulière. Le nombre important de données multimédia disponibles requièrent la création d'outils performants pour accéder rapidement à l'information. Les systèmes de recherche d'information d'images et/ou de vidéos ont été conçus pour répondre à cette forte demande. Ils doivent effectuer toutes les requêtes le plus rapidement possible avec pertinence tout en essayant de combler au mieux le fossé sémantique [Smeulders et al. (2000)] qui existe entre les descripteurs bas niveau d'une image et les concepts sémantiques (descripteurs de haut niveau). Dans ce domaine de recherche d'information multimédia [Djeraba et al. (2006)], de nombreuses techniques existent : la recherche séquentielle, les recherches basées sur des méthodes de classification [Pedrycz et al. (2008)], celles utilisant des structures d'indexation [Shen et al. (2005); Urruty et al. (2008)] ou encore les méthodes basées les retours utilisateur [Kelly et Teevan (2003)].

Motivés pour les besoins existants dans le domaine de la recherche d'information multimédia, nous proposons d'optimiser les avantages des approches ci dessus pour construire notre

propre framework. Nous proposons de combiner une technique de classification à une méthode de sélection de caractéristiques et une structure d'indexation. Ce modèle a pour avantage principal de sélectionner un sous ensemble de données avec une sélection supervisée des caractéristiques en sélectionnant les descripteurs de bas niveau les plus appropriés en fonction du thème de la recherche. L'ensemble de ce framework a donc pour objectif l'optimisation de l'efficacité de la recherche tout en gardant la meilleure pertinence possible . La suite de cet article s'organise comme suit : la section 2 décrit notre framework de recherche d'information multimédia en détaillant l'ensemble des méthodes utilisées. Les résultats expérimentaux basés sur la base de données TrecVid2008 sont présentés dans la Section 3. La section 4 conclue l'article.

2 Notre approche

Notre approche se décompose en quatre parties : la classification, la sélection de caractéristiques, l'indexation et la recherche. Les caractéristiques de bas niveau sont utilisées pour représenter les informations du contenu d'une image. Ces informations sont utiles à notre algorithme de classification pour identifier les catégories sémantique pertinentes. La deuxième étape consiste à sélectionner les caractéristiques pertinentes à la recherche pour en réduire le coût. Ensuite, une structure d'indexation multidimensionnelle détaillée dans Urruty et al. (2008) et adaptée au framework proposé et aux nombreux descripteurs bas niveau, est utilisée pour optimiser le temps de recherche. Finalement, une distance de similarité est utilisée afin de retourner les résultats les plus pertinents d'un sous ensemble de données issu d'un apprentissage supervisé et d'une combinaison pondérée des descripteurs bas niveau.

2.1 Algorithme de classification par concepts sémantiques

L'identification du contenu d'une image se fait naturellement de la vue globale à des régions plus précises. Les informations de scène extraites par la perception humaine à partir d'une vue globale, un "aperçu rapide", sont celles qui permettent à l'utilisateur d'évaluer rapidement les thèmes de l'image. Ensuite les détails de l'image permettent la compréhension totale du contenu des informations de l'image. Notre approche de classification se base sur l'hypothèse que des images proposant un contenu global similaire contiendront des thèmes sémantiques similaires. De plus, des scènes d'un même groupe ont une forte probabilité d'avoir un aspect ou une structure générale similaire avec des éléments en commun [Oliva et Torralba (2006)]. Par exemple, des images d'une ville contiendront des caractéristiques géométriques similaires à toutes constructions humaines. Les espaces verts ou les scènes de nature ont des couleurs très ressemblantes (couleurs du ciel, des arbres, des montagnes ou du désert). Nous avons donc développé un outil de classification en fonction des caractéristiques de scène basé sur les concepts globaux. Ces caractéristiques sont calculées à partir des filtres de Gabor [Howarth et Rüger (2004)]. Une fois les caractéristiques de scène extraites, nous utilisons la méthode de classification SVM [Gomez-Chova et al. (2008)] pour catégoriser les images en catégories de concept sémantique, par exemple "nature", "urbain", "nuit".

2.2 Sélection de caractéristiques de bas niveau

Basé sur les résultats de la classification, notre algorithme de sélection de caractéristiques permet de réduire le nombre de descripteurs bas niveau à utiliser pour la recherche. Nous avons choisi d'utiliser quatre descripteurs bas niveau provenant de la norme mpeg-7 : *colour structure, colour layout, homogeneous texture* et *edge histogram*. Cela représente au total 410 dimensions. Si aucune sélection n'est réalisée, le coût d'une recherche est trop élevé dû à une complexité liée au nombre de dimensions. La pertinence des descripteurs bas niveau utilisés seuls est très différente en fonction de la catégorie de concept sémantique de l'image requête. De plus, la précision des recherches peut même diminuer en utilisant certains des descripteurs bas niveau. Il est donc important de sélectionner intelligemment les meilleurs descripteurs pour effectuer la requête, de plus cela permet d'améliorer l'efficacité de celle ci. Pour notre approche, nous proposons une méthode de sélection de caractéristiques par apprentissage basée sur les catégories de concept sémantique. Nous utilisons l'algorithme SVM sur une base de données d'apprentissage afin d'évaluer le potentiel de chaque descripteur pour chaque classe. Cette évaluation permet d'affecter des poids aux descripteurs bas niveau pour les futures recherches. Pour cela, nous avons calculé la mesure F1 de chaque classe de notre base d'apprentissage en fonction de chaque descripteur avec SVM. La mesure F1 est la moyenne harmonique entre la précision et le rappel.

2.3 Méthodologie de recherche

La recherche ou l'accès à l'information contenue dans notre base de données se fait en deux étapes. Tout d'abord, pour chaque image requête, nous déterminons les catégories de concept sémantique auxquelles l'image est la plus proche grâce à l'utilisation des filtres de Gabor et du résultat de la méthode de classification SVM comme détaillé dans une section précédente. Par conséquent, il est possible d'estimer les descripteurs bas niveau les plus pertinents pour effectuer la recherche (voir 2.2).

Soit un ensemble d'images requête d'un scénario de recherche, la première étape de notre méthodologie est celle de la classification des images requête en une catégorie de concept. On note Ω, l'ensemble des classes. On dénote k, le nombre de catégories sélectionnées, où $k < card(\Omega)$. On dénote aussi S^k_Ω, un sous ensemble d'images de la classe k dans lequel la requête s'effectue.

Soit $R(d, S^k_\Omega)$, la pertinence d'un descripteur bas niveau d en fonction du sous ensemble S^k_Ω, cette pertinence est calculée grâce à la valeur de la mesure $F1_c$ de chaque classe c obtenue en phase d'apprentissage (voir la Section 2.2) par la formule suivante :

$$R(d, S^k_\Omega) \; - \; (\sum_{c=0}^{k} F1_c(d)) \qquad (1)$$

L'évaluation des paramètres α et k est introduite dans la section suivante.

Pour chaque descripteur bas niveau d d'une requête image, notre algorithme retourne 1000 résultats. Pour cela, notre modèle de recherche effectue une recherche par descripteur pour obtenir $i \times d \times 1000$ résultats de recherche incluant des possibles doublons. Un algorithme de fusion permet de réduire le nombre de résultats en sélectionnant les 1000 premiers en fonction du nombre d'occurrences de chaque résultat puis de leur distance à la requête image.

3 Expérimentations

Nos expérimentations utilisent sur la base de données TrecVid2008 [Smeaton et al. (2006)] disponible en ligne. TrecVid2008 est une base de données de 35.000 séquences vidéo représentées par 730.000 images clé. Il existe 48 différents sujets de recherche avec un ensemble d'images requête. En fonction de la sémantique de ces sujets, nous définissons un ensemble de sept catégories de concept sémantique : "ville", "humain", "nature", "intérieur", "extérieur", "nuit", et "véhicule". La base de données que nous avons choisie pour notre apprentissage est la base de données de TrecVid2007. 700 images par catégorie de concept sémantique sont sélectionnées pour lancer l'apprentissage. Les résultats de recherche sont obtenus à partir d'une recherche sur les 730.000 images et les performances de notre approche sont analysées par l'outil d'évaluation *TrecEval Tool*. Pour ces expérimentations, nous utilisons quatre descripteurs bas niveau Mpeg-7 : *Colour Layout*, *Colour Structure*, *Homogenous Texture* et *Edge Histogram*. Nous comparons les performances de notre approche aux performances obtenues par la recherche séquentielle utilisant l'ensemble des descripteurs.

Nous proposons dans nos expérimentations d'étudier la précision sur les 100 premiers résultats, le nombre total de résultats pertinents retournés et le temps de réponse à une requête en jouant sur les paramètres suivants : le nombre k de classes et le nombre α de descripteurs bas niveau à utiliser pour la recherche. Comme précisé précédemment, nos *résultats de référence* sont les résultats obtenus par la recherche séquentielle sur l'ensemble de la base de données, i.e. l'ensemble des classes, et aussi en utilisant l'ensemble des descripteurs (ce qui signifie $\alpha = 4$ et $k = 7$).

(α, k)	P(100)	Nb pertinents	Temps (s)
Référence	0.15	1055	11.38
(1, 2)	0.09 (-36.8%)	744 (-29.5%)	0.7 (-93.8%)
(2, 2)	0.13 (-10.7%)	926 (-12.2%)	2.3 (-80%)
(3, 4)	0.15 (-1.1%)	1002 (-5%)	4.5 (-60%)
(4, 4)	0.16 (+8.3%)	1107 (+4.9%)	6.9 (-39%)
(4, 7)	0.15 (-0%)	1055 (-0%)	9.1 (-20%)

TAB. 1 – *Sélection de différentes valeurs pour α et k*

Le tableau 1 présente les meilleures performances obtenues pour diverses valeurs de α et k. En effet, si on sélectionne 4 classes et les $\alpha = 4$ descripteurs bas niveau pour la recherche, nous obtenons les meilleures performances en terme de précision et du nombre de résultats pertinents obtenus. Cependant le temps de réponse est de 6,9 secondes ce qui n'est pas envisageable pour un moteur de recherche en ligne. L'utilisation de $k = 2$ classes et des $\alpha = 2$ meilleurs descripteurs permet d'avoir de bons résultats en général. En effet, faire ce choix diminue les performances de la recherche de 12% en précision et en nombre de résultats pertinents par rapport aux résultats de référence, et de 20% par rapport aux meilleurs résultats, mais diminue aussi le temps de réponse de 80%. Nous avons aussi observé que l'utilisation de notre structure d'indexation permet de gagner 20% sur le temps de réponse par rapport au temps de la recherche séquentielle sur l'ensemble de données. Notre approche propose donc un très bon compromis entre précision et efficacité.

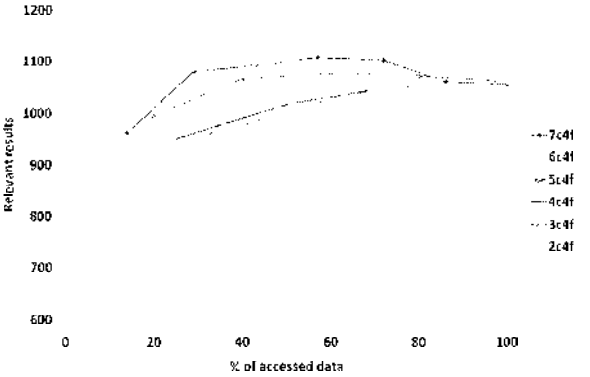

FIG. 1 *Nombre de résultats pertinents en fonction du pourcentage de données accédées pour différent nombres de classes lors de l'apprentissage*

A l'issue de l'ensemble de ces expérimentations et résultats, une nouvelle question se soulève : "Quelle est l'effet du nombre de classes utilisées lors de l'étape d'apprentissage pour la classification sur l'efficacité de la recherche ?" Pour répondre à cette question, nous avons lancé de nouvelles expérimentations en faisant varier le nombre de classes, de 2 à 7, pour l'apprentissage lors de l'étape de la classification. Afin d'obtenir des résultats cohérents, nous avons transformé le nombre de classes utilisées lors de la recherche par le pourcentage approximatif de données accédées. Nous assumons qu'en moyenne sur l'ensemble des requêtes, le nombre de données dans les classes sélectionnées est similaire. Par exemple, 33% dans la ligne 3c4f signifie que l'algorithme a sélectionné une classe sur les trois disponibles pour faire la recherche, 66%, 2 classes et 100% les trois classes, donc l'ensemble des données.

La figure 1 présente le nombre de résultats pertinents en fonction du pourcentage de données accédées. Nous observons que si la recherche accède à plus de 60% des données, les résultats sont identiques. En dessous de 50% de données accédées, utiliser 6 ou 7 classes lors de l'apprentissage permet une meilleure précision dans la recherche. Le temps de réponse étant linéaire au pourcentage de données accédées, nous concluons qu'utiliser un sous ensemble de classes avec un apprentissage de 6 ou 7 classes pour la classification permet d'obtenir les meilleures performances et un bon compromis temps de réponse / précision.

4 Conclusion

Cet article propose une discussion sur la possibilité de combiner une méthode de classification, de sélection de caractéristiques et d'indexation pour une application tournée vers la recherche de documents multimédia. L'avantage d'une telle approche est de fusionner la recherche d'information utilisant les descripteurs bas niveau à une classification en catégories de concept sémantique. La plupart des méthodes existantes ne proposent pas ce compromis entre les concepts de haut niveau et la recherche de bas niveau.

Pour résumer, l'approche proposée s'intéresse à une combinaison peu étudiée à ce jour entre la recherche bas niveau et haut niveau. Elle a démontré des résultats prometteurs avec un potentiel justifiant la nécessité d'un investissement plus important pour de futures recherches.

Références

Djeraba, C., M. Gabbouj, et P. Bouthemy (2006). Multimedia indexing and retrieval : ever great challenges. *Multimedia Tools Appl. 30*(3), 221–228.

Gomez-Chova, L., G. Camps-Valls, J. Munoz-Mari, et J. Calpe (2008). Semisupervised image classification with laplacian support vector machines. *Geoscience and Remote Sensing Letters, IEEE 5*(3), 336–340.

Howarth, P. et S. M. Rüger (2004). Evaluation of texture features for content-based image retrieval. In *CIVR'04*, Volume 3115 of *LNICS*, pp. 326–334.

Kelly, D. et J. Teevan (2003). Implicit feedback for inferring user preference : a bibliography. *SIGIR Forum 37*(2), 18–28.

Oliva, A. et A. Torralba (2006). Building the gist of a scene : the role of global image features in recognition. *Progress in brain research 155*, 23–36.

Pedrycz, W., A. Amato, V. Di Lecce, et V. Piuri (2008). Fuzzy clustering with partial supervision in organization and classification of digital images. *Fuzzy Systems, IEEE Transactions on 16*(4), 1008–1026.

Shen, H. T., B. C. Ooi, et X. Zhou (2005). Towards effective indexing for very large video sequence database. In *SIGMOD '05*, NY, USA, pp. 730–741.

Smeaton, A. F., P. Over, et W. Kraaij (2006). Evaluation campaigns and trecvid. In *MIR '06 workshop*, NY, USA, pp. 321–330.

Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta, et R. Jain (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell. 22*(12).

Urruty, T., C. Djeraba, et J. M. Jose (2008). An efficient indexing structure for multimedia data. In *Proceedings of ACM MIR 08*, Vancouver, Canada. ACM.

Summary

As content-based multimedia applications become increasingly important, demand for technologies on Content Based Image and Video Retrieval (CBIVR) is growing, where the retrieval results are expected to be in line with the query examples in semantic meanings. Based on extensive literature survey on existing CBIVR algorithms, we propose in this paper an integrated approach for concept based image and video retrieval, where scene information in the images is taken into account in exploiting the image concept characters together with a feature selection model to reduce the feature dimension and determine the best low level visual features for each query. In comparison with existing retrieval models, our contribution can be highlighted as: (i) a concept based classification algorithm; (ii) a feature selection model; (iii) a multidimensional indexing structure. The efficiency and the effectiveness of the proposed framework have been established with an extensive experimentation on the TRECVid2008 data collection.