Annotation d'images sur de grands corpus réels de données

Limites des histogrammes de couleurs

Pierre Tirilly* — **Vincent Claveau*** — **Patrick Gros****

* *IRISA / CNRS*
Campus de Beaulieu
35042 Rennes, France
{Pierre.Tirilly,
Vincent.Claveau}@irisa.fr

** *IRISA / INRIA*
Campus de Beaulieu
35042 Rennes, France
Patrick.Gros@irisa.fr

RÉSUMÉ. *Dans cet article, nous vérifions les limitations des techniques d'annotation d'images sur un corpus réel et de grande taille. Pour cela, nous utilisons un corpus de documents texte-images composé de plus de 25000 articles de presse, sur lequel nous évaluons la similarité entre une recherche basée sur le texte et une recherche basée sur l'image. Les systèmes de recherche que nous utilisons sont des outils communs de la Recherche d'Information (RI). Les résultats montrent que les systèmes de RI utilisés donnent des résultats différents : l'association simple des descripteurs utilisés ici ne semble donc pas appropriée pour l'annotation de données réelles. Cela met en avant la nécessité de développer un modèle permettant de prendre en compte, de manière cohérente, des descripteurs textuels et visuels pour réaliser l'indexation sémantique d'images.*

ABSTRACT. *In this paper, we check the limitations of image annotation on a large real corpus. We built a corpus of documents containing text and pictures, using more than 25000 press articles. We use this data to compare the similarity between a text-based retrieval and an image-based retrieval. The retrieval systems we use are common Information Retrieval (IR) tools. Results show that the two IR systems we used do not work well together, so associating the descriptors we used is not suitable to annotate real data. Then it is necessary to find a model than can handle textual and visual descriptors together, in order to make image annotation possible.*

MOTS-CLÉS : *Annotation d'images, propagation de mots-clefs, indexation texte-image, fossé sémantique, histogramme de couleurs*

KEYWORDS: *Image annotation, keywords propagation, text-image indexing, semantic gap, color histogram*

CORIA 2007, pp. 473–478
Saint-Étienne, 28–30 mars 2007

1. Introduction

Le développement et la démocratisation des outils informatiques ont provoqué une véritable explosion du nombre de documents numériques. Se pose alors le problème, pour les utilisateurs, du stockage et de l'accès à ces données : comment retrouver les documents souhaités à partir d'une requête en langage naturel ? Dans le cas des documents textuels, l'appariement entre les mots-clefs de la requête et le contenu du document est naturel, et les méthodes de Recherche d'Information (RI) permettent aujourd'hui d'indexer et de rechercher les documents de ce type. Dans le cas des images, les techniques actuelles de RI sont dites « par le contenu », *i.e.* les images sont décrites par leur caractéristiques visuelles (couleur, texture, forme…) uniquement. Avec de tels descripteurs, il n'est pas possible d'extraire des concepts des images, que l'on pourrait rapprocher des concepts contenus dans les mots des requêtes. Ce problème, appelé « fossé sémantique », est caractéristique de l'indexation de documents multimédias : il est difficile d'associer des descripteurs « sémantiques », c'est-à-dire identifiant des concepts, à des documents représentés par des descripteurs numériques.

Dans les approches actuelles, la tâche d'indexation sémantique des images revient à associer à une image des mots-clefs décrivant le sens porté par l'image : on parle alors d'annotation (semi-)automatique d'images. Par exemple, les techniques de propagation de mots-clefs proposent d'étendre les annotations d'images déjà annotées à de nouvelles images en se basant sur les similarités visuelles entre ces images. Dans cet article[1], nous proposons de vérifier les limites des approches s'appuyant sur des histogrammes de couleurs pour l'annotation de corpus réels et de grande taille, en se basant sur des outils éprouvés de la RI.

2. Travaux connexes et approche proposée

Plusieurs modèles d'annotation d'images par des mots-clefs ont été proposés dans la littérature. On peut distinguer deux grandes approches (Hare *et al.,* 2006). Les approches dites *top-down* sont basées sur des connaissances *a priori* comme des ontologies ou des dictionnaires (Schreiber *et al.,* 2001, Kompatsiaris *et al.,* 2004). Par opposition, les approches dites *bottom-up* utilisent les données pour en extraire les relations existant entre mots et images. Parmi ces modèles on peut citer les modèles génératifs ou probabilistes (Barnard *et al.,* 2003, Jeon *et al.,* 2003), les modèles basés sur les co-occurences (Mori *et al.,* 1999), ou les modèles discriminants, basés sur des classifieurs (Jing *et al.,* 2004, Chang *et al.,* 2003).

On peut émettre plusieurs reproches envers ces modèles et la manière dont ils sont validés. D'une part, ils nécessitent des connaissances *a priori* sur les données. Ainsi même les modèles *bottom-up*, bien que s'appuyant en principe sur les données, utilisent des hypothèses *a priori* sur celles-ci, telles que leur distribution statistique ou

1. Ce travail a été réalisé avec le soutien du réseau d'excellence européen MUSCLE du 6e P.C.R.D.T., de la région Bretagne et du C.N.R.S.

le nombre de catégories à différencier. D'autre part, les données utilisées pour valider ces modèles sont rarement des corpus de données issus d'applications réelles, mais des banques d'images choisies spécifiquement et dont les catégories sous-jacentes sont connues et clairement définies. En particulier, il a été montré que la base d'images COREL, très prisée en annotation automatique, peut donner, pour un même modèle, des résultats très variables et donc peu généralisables (Müller *et al.,* 2002). Enfin, d'une manière plus générale, l'association simple de mots-clefs et de descripteurs visuels ne paraît pas suffisante pour décrire le sens des images : des descripteurs numériques, comme par exemple les histogrammes de couleurs, ne semblent pas pouvoir représenter les concepts contenus dans une image (par exemple une voiture) tels qu'ils le sont par des mots.

Dans cet article, nous proposons de vérifier les limitations de ces approches d'annotation sur de grands corpus réels. Pour cela, nous avons constitué un corpus de documents bimodaux (textes et images) composé d'articles de presse. Nous disposons donc de données correspondant à un problème réel : l'indexation d'images de presse. Sur un tel corpus, il n'existe pas de classification *a priori* des documents. De plus, ce corpus est généraliste, aussi bien en termes d'images que de vocabulaire, ces articles traitant de sujets variés (politique, sports, culture…). Enfin, chaque image de ce corpus est accompagnée d'une légende la décrivant, comme les légendes utilisées par les archivistes. Ce sont ces légendes qui sont utilisées, dans notre expérience, comme partie textuelle des documents.

Sur la base de ce corpus, nous avons mis en place une expérience qui vérifie l'adéquation des outils de recherche d'information textuels et visuels : disposant d'un système de RI textuelle et d'un système de RI visuelle, nous cherchons à évaluer la corrélation entre les listes de résultats obtenues par chacun des systèmes. S'il existe une corrélation entre ces résultats, alors ces systèmes identifient les concepts contenus dans le texte et dans l'image de la même manière ; utiliser la propagation de mots-clefs sur la base de tels systèmes serait donc légitime. Inversement, s'il n'y a aucune corrélation entre les résultats des systèmes de RI utilisés, cela signifie que ces systèmes identifient des concepts différents. Cela mettrait donc en avant une limitation des systèmes d'annotation, qui mettent en relation mots-clefs et descripteurs numériques alors que la similarité entre mots-clefs est indépendante de la similarité entre descripteurs numériques.

Les systèmes de RI retenus pour l'expérience sont des outils utilisés communément en RI, parmi les plus au point de l'état de l'art :

– Pour la RI textuelle, nous utilisons le système LEMUR, paramétré de façon à se comporter comme le système OKAPI (Robertson *et al.,* 1996) : la partie textuelle de chaque document (dans notre cas, les légendes des images) est représentée par plusieurs mots-clefs extraits automatiquement du corpus. Pour chaque document, l'importance accordée à chaque mot-clef est calculée à partir de la formule de pondération $BM25$ avec les paramètres par défaut. À l'issue d'une recherche, les documents sont classés selon leur probabilité d'être pertinents pour la requête donnée.

– Pour la RI visuelle, nous avons utilisé des histogrammes de couleurs. Nous avons choisi des histogrammes à cumul additif et pondérés par le Laplacien, dans l'espace colorimétrique RGB. Ces histogrammes sont obtenus en cumulant des histogrammes locaux calculés pour chaque partie d'une image divisée en quatre zones. Ils permettent de tenir compte de la répartition spatiale des couleurs, et offrent de bonnes performances (Boujemaa *et al.,* 2002). Pour calculer la similarité entre les histogrammes, nous avons testé deux distances classiques : L_1 et L_2, cas particuliers de la distance de Minkowski L_p : $L_p(q,r) = \left(\sum_{i=1}^{n} |q_i - r_i|^p \right)^{1/p}, \quad \forall\, q,\, r \in \mathbb{R}^n$

3. Expérience

3.1. *Données*

Les données utilisées pour cette expérience sont des articles de presse téléchargés sur le site www.tv5.org. Les articles contiennent un texte (corps de l'article) et une ou plusieurs images accompagnées de leur légende. Pour l'expérience, nous utilisons uniquement les légendes comme composante textuelle des documents. L'indexation textuelle est réalisée classiquement, grâce à des mots-clefs extraits automatiquement des légendes. Le corpus de l'expérience est composé de 25753 images et des légendes associées.

3.2. *Protocole*

Le protocole employé est le suivant :

1) choisir aléatoirement n documents texte-image en tant que requêtes. Pour obtenir des résultats significatifs, nous avons utilisé $n = 200$ requêtes.

2) indexer le corpus de légendes à l'aide de LEMUR.

3) calculer les histogrammes de couleurs de chaque image.

4) pour chaque document requête d_r, constitué d'une image et de la légende l'accompagnant :

a) effectuer une recherche sur le texte de la légende avec LEMUR ; on obtient une liste l_1 de documents résultats ordonnés par pertinence.

b) effectuer une recherche sur les images selon la distance choisie (L_1 ou L_2). On obtient une liste l_2 de résultats ordonnés selon leur distance à la requête.

c) calculer la corrélation des listes l_1 et l_2. En RI, les résultats pertinents se trouve en tête de classement, nous avons donc calculé les corrélations sur des listes tronquées aux 10, 20, 50, 100, 500 et 1000 premiers résultats. Le coefficient utilisé pour calculer les corrélations est le τ de Kendall modifié avec une pénalité $p = 1$, noté $K^{(1)}$. Cette mesure permet d'estimer la corrélation entre deux listes ne partageant pas tous leurs éléments (Fagin *et al.,* 2003, pour la formule de $K^{(1)}$). Une valeur de $K^{(1)} = 0$ indique une corrélation totale entre les listes, et une valeur de $K^{(1)} = 1$ une absence de corrélation.

4. Résultats

Le tableau 1 résume les résultats obtenus sur les 200 requêtes. On remarque que, globalement, les valeurs de $K^{(1)}$ sont plutôt proches de 1, ce qui indique une absence de corrélation entre les listes issues du système de RI textuel et celles issues du système de RI visuel. De plus, les écarts-types diminuent avec la taille des listes considérées. Ainsi, les résultats obtenus par requête sont homogènes pour les grandes listes, et instables pour les petites listes. Il n'y a donc pas de cohérence entre les résultats délivrés par le système de RI textuel et ceux délivrés par le système de RI visuel. De plus, il y a une forte corrélation (coefficient de Spearman moyen $\rho = 0.94$) entre les listes de résultats obtenues avec L_1 et L_2. Ces deux distances sont donc proches, ce qui explique leurs résultats similaires.

k	10	20	50	100	500	1000
L_1	0.63 (0.21)	0.77 (0.19)	0.88 (0.12)	0.93 (0.09)	0.96 (0.06)	0.95 (0.04)
L_2	0.64 (0.21)	0.77 (0.18)	0.88 (0.12)	0.93 (0.09)	0.96 (0.06)	0.95 (0.04)

Figure 1. *τ de Kendall modifié $K^{(1)}$ moyen (et écarts-types) sur les listes tronquées à k éléments, pour chaque distance testée.*

Ces résultats montrent une faiblesse dans les processus d'annotation. Cela ne signifie pas nécessairement que l'annotation d'images est impossible à réaliser, mais qu'elle soulève des difficultés non révélées par les travaux sur des bases trop spécifiques. D'une part, cela montre la nécessité de trouver un modèle permettant une utilisation conjointe et cohérente des descripteurs textuels et visuels. D'autre part, les systèmes de RI utilisés pour l'expérience ont une influence importante sur les résultats. Le choix d'autres systèmes ou d'autres descripteurs pourrait changer ces résultats. Plus généralement, le choix des bons systèmes de RI et des bons descripteurs est essentiel pour réaliser l'annotation automatique d'images.

5. Conclusion et perspectives

Nous avons testé des limitations de l'annotation d'images sur un grand corpus réel de documents texte-images. Pour cela, nous avons mis en place une expérience qui confronte les résultats de deux systèmes de RI : l'un textuel (OKAPI), l'autre visuel (histogrammes de couleurs). Les résultats de cette expérience montrent qu'il n'existe aucune corrélation entre les classements effectués par chacun des systèmes pour une requête donnée. Cela met en évidence un manque de cohérence entre les systèmes de RI textuels et visuels actuels basés sur les histogrammes de couleurs, et donc la nécessité de trouver un modèle pour mettre efficacement en relation les descripteurs textuels et numériques.

Plusieurs axes peuvent être explorés pour réaliser l'annotation d'images à partir de documents mêlant texte et images. Tout d'abord, on peut envisager une sélection des informations pertinentes au sein du texte : en isolant les parties du texte relatives à l'image, on pourrait en extraire les mots-clefs décrivant au mieux le contenu sémantique de l'image. De plus, il est possible d'utiliser des descripteurs de plus haut

niveau, pour mieux annoter les images. Du point de vue du texte, les techniques de Traitement Automatique du Langage (TAL) permettent d'extraire des termes spécifiques (par exemple les entités nommées) ou des relations entre termes (par exemple la synonymie) qui peuvent être exploités dans le cadre de l'annotation d'images. Similairement, dans le cas des descripteurs visuels, le procédé d'annotation pourrait profiter de l'apport d'autres descripteurs comme la texture, ou de descripteurs de plus haut niveau, capables d'extraire des informations visuelles plus sémantiques, par exemple un détecteur de visages. Enfin, de nombreuses techniques, issues d'autres problèmes, peuvent s'adapter pour unifier les descripteurs visuels et textuels (analyse de la sémantique latente, classifieurs, *clustering*...).

6. Bibliographie

Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D. M., Jordan M. I., « Matching words and pictures », *Journal of Machine Learning Research*, vol. 3, p. 1107-1135, 2003.

Boujemaa N., Boughorbel S., Vertan C., « Description de la répartition spatiale de la couleur pour l'Indexation d'Images », *RFIA02 : 13e Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, 2002.

Chang E., Kingshy G., Sychay G., Wu G., « CBSA : content-based soft annotation for multimodal image retrieval using Bayes point machines », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, p. 26-38, 2003.

Fagin R., Kumar R., Sivakumar D., « Comparing top k lists », *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2003.

Hare J., Sinclair P., Lewis P. H., Martinez K., Enser P., Sandom C., « Bridging the Semantic Gap in Multimedia Information Retrieval : Top-down and Bottom-up approaches », *Proceedings of Mastering the Gap : From Information Extraction to Semantic Representation, 3rd European Semantic Web Conference*, 2006.

Jeon J., Lavrenko V., Manmatha R., « Automatic image annotation and retrieval using cross-media relevance models », *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003.

Jing F., Li M., Zhang H.-J., Zhang B., « Keyword Propagation for image Retrieval », *ISCAS '04 : Proceedings of the 2004 International Symposium on Circuits and Systems*, 2004.

Kompatsiaris I., Avrithis Y., Hobson P., Strintzis M. G., « Integrating Knowledge, Semantics and Content for User-Centred Intelligent Media Services : The aceMedia Project », *WIAMIS '04 : Proceedings of Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

Mori Y., Takahashi H., Oka R., « Image-to-word transformation based on dividing and vector quantizing images with words », *MISRM '99 : Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

Müller H., Marchand-Maillet S., Pun T., « The Truth about Corel - Evaluation in Image Retrieval », *CIVR '02 : Proceedings of the International Conference on Image and Video Retrieval*, 2002.

Robertson S. E., Walker S., Hancock-Beaulieu M., Gull A., Lau M., « Okapi at TREC-4 », *TREC-4 : Proceedings of the 4th Text REtrieval Conference*, 1996.

Schreiber A. T., Dubbeldam S., Wielemaker J., Wielinga B., « Ontology-based photo annotation », *IEEE Intelligent Systems*, vol. 16, p. 66-74, 2001.