Fusion de multi-modalités et réduction par sémantique latente Application à la recherche de documents multimédia et à l'annotation automatique d'images

Trong-Ton Pham, Jean-Pierre Chevallet, Joo-Hwee Lim

Fusion de multi-modalités et réduction par sémantique latente

Application à la recherche de documents multimédias et à l'annotation automatique d'images

Pham Trong Ton, Jean-Pierre Chevallet, Lim Joo Hwee

Laboratoire Image Perception Access & Language (IPAL) - UMI CNRS 2955
21 Heng Mui Keng Terrace, 119613, Singapore
{ttpham,viscjp,joohwee}@i2r.a-star.edu.sg

RÉSUMÉ. *Ce papier étudie la "sémantique latente" entre des éléments visuels et textuels d'une collection multimédia, appliquée à deux tâches : (1) la Recherche de Document Multimédia (RDM) contenant des images et du texte ; et (2) l'Annotation Automatique d'Images (AAI). La sémantique latente, habituellement utilisée dans l'indexation textuelle, est mise à profit ici pour faire apparaître des liens entre les descriptions textuelles et visuelles des images. Nous avons ainsi deux contributions principales. Il s'agit d'une part, de la première étude sur l'influence de la sémantique latente entre termes textuels et visuels, sur une grande collection de documents. En effet, cette méthode est testée sur une collection de 20000 images touristiques. D'autre part, nous démontrons que la fusion des différents modalités d'image (i.e. termes visuels vs textuels, et différentes méthode de représentations d'image) améliore le résultat d'une annotation automatique des images par du texte. Nos collections de test sont la base d'images annotées de COREL et la base d'ImageCLEF 2006.*

ABSTRACT. *This paper studies Latent Semantic Analysis (LSA) between visual and textual elements of a multimedia collection, applied on two different tasks: (1) multimedia document retrieval (MDR) and (2) automatic image annotation (AIA). Latent semantics, usually used for text indexing, is applied to discover hidden semantic relations between visual and textual descriptors. The contributions of this paper are twofold. First, to the best of our knowledge, this work is the first study of the influence of LSA on the retrieval of a significant number of multimedia documents (i.e. collection of 20000 tourist images). Second, it shows how different image representations (region-based and keypoint-based) can be combined by LSA to improve automatic image annotation. The document collections used for these experiments are the Corel photo collection and ImageCLEF 2006 collection.*

MOTS-CLÉS : *Recherche d'information, annotation automatique, sémantique latente, modèle de fusion*

KEYWORDS: *Information retrieval, automatic image annotation, latent semantic, multimedia fusion*

1. Introduction

La recherche d'images par le contenu visuel est une tâche difficile à automatiser. Plusieurs méthodes ont été proposées ces dernières années pour construire un système de *Recherche d'Images par le Contenu*[1] [SME 00]. Cependant, le manque d'informations explicites sur la demande de l'utilisateur, et les réelles difficultés pour un ordinateur d'interpréter effectivement le contenu d'une image, font que ce problème d'indexation et de recherche d'images est particulièrement difficile.

En fait, la recherche d'image par le contenu se fait sur le *signifié* de l'image, c'est à dire sur son interprétation par un lecteur humain, ainsi que par rapport au contexte de sa capture. Ce contexte doit être considéré comme une méta information. Il s'agit par exemple, de rechercher une image par le nom de l'endroit visité, le nom d'un l'animal photographié, le nom du personnage en gros plan etc... Pourtant, les techniques actuelles de vision par ordinateur ne nous permettent que d'extraire des images, des *caractéristiques visuelles de bas niveau* (couleur, texture ou point d'intérêt). La distance à parcourir pour interpréter automatiquement le contenu d'une image semble encore très importante. La question récurrente est donc posée en ces termes : comment utiliser au mieux les caractéristiques visuelles de bas niveau pour les relier automatiquement à des concepts ? Ce problème est connu sous le terme "semantic gap" ou *fossé sémantique* [ZHA 02].

Lorsque les modalités textuelles et visuelles sont réunies dans un même document, il parait judicieux d'exploiter simultanément ces deux types de contenu. Sachant, d'une part qu'il est plus facile d'associer automatiquement du sens à un texte qu'à une image, et que d'autre part, il est plus facile de comparer des images similaires, on peut espérer qu'une complémentarité entre l'image et le texte soit propice à une meilleure indexation, par rapport aux deux médias pris séparément. Cependant, faire fonctionner un système combinant ces médias dans une *Recherche de Documents Multimédias* (RDM) est un problème loin d'être évident à résoudre.

La solution la plus immédiate est un *système de recherche d'image par le texte*[2]. Actuellement, la plupart des systèmes de recherche d'images accessibles au grand public (Google Images, Yahoo !, Flickr ...) se basent sur des informations provenant d'annotations de l'image. Par exemple, *Google* indexe les images du web en fonction du texte qui les entoure (nom du fichier, description, lien du web...) et *Flickr* indexe les images de sa base de données en fonction des mots-clés ("tags", location, catégories...) que les utilisateurs attribuent eux-même aux images. Cette approche est naturelle : le texte est plus ou moins régie par des règles de grammaires, et par des règles sémantiques que l'on peut décrire et faire valider par une machine. En revanche, il est beaucoup plus difficile de construire une *grammaire pour les images*. Également, alors qu'il existe des ontologies de concepts associées à des lexiques de termes, il existe très peu d'ontologies visuelles.

---

1. Content-Based Image Retrieval (CBIR)
2. Annotation-Based Image Retrieval (ABIR)

Indexer le texte associé à une image, plutôt que l'image elle même est donc une solution techniquement plus facile à mettre en œuvre. Il vient alors l'idée d'associer automatiquement à de nouvelles images, des annotations existant dans une base [INO 04], selon l'hypothèse (discutable), que des images similaires visuellement doivent partager des annotations textuelles [CHE 06]. Dans la littérature, ces systèmes sont appelés *Annotation Automatique d'Images*(AAI) ou *Automatic Image Tagging* en anglais.

Influencé par la recherche en apprentissage automatique, il existe deux approches principales pour le problème d'annotation et de recherche d'images. Tout d'abord, une première approche est basée sur un apprentissage supervisé. Des images d'entraînement sont classées manuellement. Par exemple, une classe est définie pour chaque mot-clé ou pour un ensemble de mots-clés. Ensuite un classifieur binaire est entraîné pour chaque classe : c'est la phase d'apprentissage. Lors de la phase de classification, une nouvelle image est présentée successivement à toutes les classifieurs. La classification positive d'un classifieur revient à reconnaître l'image comme visuellement similaire aux exemples positifs et permet de lui associer le mot clé correspondant.

Une autre approche consiste à découvrir automatiquement les liens cachés entre les éléments visuels et les annotations textuelles en utilisant des méthodes apprentissage non-supervisé [LI 03, FEN 04, LAV 03]. La non supervision permet de se passer de la phase d'apprentissage et de la classification manuelle. Pour cela, cette technique introduit un ensemble de *variables latentes* sensées représenter la co-distribution des éléments visuels et textuels. Étant donné une nouvelle image sans annotation, les caractéristiques visuelles sont extraites et la fonction de similarité probabiliste va retourner l'état qui maximise la densité probabiliste de l'annotation textuelle et l'élément visuel. Finalement, les annotations sont triées par les valeurs de probabilités.

Dans le domaine textuel, *l'analyse par sémantique latente*[3] [LAN 98] est une méthode statistique permettant de découvrir les liens entre les mots. Cette technique construit un nouvel espace d'indexation, aux dimensions plus réduites que l'espace d'origine, en groupant les dimensions, redondante, donc les termes redondants. Les dimensions d'origine représentent l'importance des termes dans les documents : la LSA fait donc l'hypothèse que la co-occurrence des termes cache une relation sémantique latente.

Pourquoi alors ne pas étendre cette idée aux images annotés ?

C'est chose fait dans Quelhas et al. [QUE 05]. Ils ont appliqué cette méthode en se basant purement sur des caractéristiques visuelles. Cependant, nous avons constaté qu'il n'existe aucun travail mettant en œuvre cette technique de LSA sur une collection mixte (texte et image) de taille significative : dans [ZHA 02], les auteurs ont mésuré l'influence de la LSA sur une collection de 20 documents seulement. Plus récemment, Monay et al. [MON 03] ont étendu l'expérimentation sur une collection de

3. Latent Semantic Analysis (LSA)

8000 images. Ils ont ainsi montré, par leurs résultats encourageants, que l'utilisation de LSA entre caractéristiques visuelle et textuelle, est une voie intéressante à explorer.

Dans le domaine de vision par ordinateur et la recherche l'information (RI), beaucoup de techniques de représentation d'une images ont été étudiées. On peut citer par exemple la segmentation par pavages rectangulaires [MOR 99], ou la segmentation en région "Blobword" [DUY 02], ou encore la détection de points d'intérêts [LOW 99]. Dans [QUE 05], Quelhas et al. ont fait une étude intéressante sur l'influence de la LSA sur la représentation d'une scène avec points d'intérêts. Nous constatons donc d'une part, que cette idée d'utiliser la technique LSA à la fois sur les éléments visuels et textuels semble prometteuse, et que d'autre part, elle mérite de plus importantes expérimentations pour mieux comprendre les effets du LSA sur la combinaison entre annotations et caractéristiques visuelles. Cet article vise donc à combler ce manque.

Nous proposons donc d'examiner plus en détail cette technique de réduction de dimension par LSA sur une collections multimedia de la manière suivante : d'une part dans la tâche de Recherche de Documents Multimédias (RDM) et d'autre part pour la tâche d'Annotation Automatique d'Images (AAI). Notre contribution au domaine est la suivante : tout d'abord, nous mesurerons l'influence réciproque et combinée des informations textuelles et visuelles extraites de documents contenant les deux modalités textuelle et visuelle ; puis nous évaluons l'amélioration apportée par le LSA dans la combinaison des deux sources d'information pour l'indexation et recherche des documents sur des grandes collections d'images (i.e. 5000 images de COREL et 20000 images d'ImageCLEF[4]) ; finalement, nous montrons que l'influence de la fusion de plusieurs représentations d'images (i.e. région et point d'intérêt) peut améliorer les résultats.

Nous présentons dans la section suivante le traitement du texte et d'image pour l'indexation les documents multimédias. En suite, nous proposons notre modèle de fusion par LSA pour les deux tâches RDM et AAI. La section 4 présente les résultats expérimentaux obtenus.

2. Traitement du texte et d'image

2.1. *Caractéristiques visuelles*

Nous présentons dans cette partie les techniques que nous proposons pour extraire les *termes visuels* qui servent à représenter les images dans un espace vectoriel. Ces termes visuels peuvent être des régions ou des point d'intérêts (voir figure 1). Une région est obtenu de deux manières différentes : (1) soit par une partition en zones rectangulaires (par exemple division d'une images en 4x4 rectangles) [MOR 99] (2) soit une segmentation en région [DUY 02] en utilisant l'algorithme de *Mean-shift* [COM 02]. Les point d'intérêts sont localisés par détection des points d'extrema par

4. http ://ir.shef.ac.uk/imageclef/

filtrage de l'image avec un opérateur de différence de gaussien[5]. Un vecteur de caractéristiques SIFT[6] [LOW 99] de taille 128 dimensions est calculé pour chaque point d'intérêt.

Figure 1. *Différentes représentations d'une image : (a) segmentation par pavage rectangulaires, (b) segmentation en régions, et (c) points d'intérêts*

Pour chaque région, nous extrayons trois caractéristiques : un histogramme de couleurs, des coefficients de Gabor, et la position du centre de la région. Nous utilisons un histogramme de 4 canaux contenant l'information R (rouge), G (vert), B (bleu) et L (luminance)[7]. Le tableau suivant récapitule les informations sur les différentes caractéristiques testées.

Tableau 1. *Les caractéristiques visuelles utilisées dans nos expérimentations*

Caractéristique	Quantification	Dimension
SIFT	(4×4) rectangles $\times$ 8 orientations	128
RGBL	16 éléments $\times$ 4 canaux	64
Gabor	6 orientations $\times$ 5 échelles	30
Location	(x_c, y_c)	2

2.2. *Le vocabulaire visuel*

Le vocabulaire visuel est le pendant des termes extraits des documents textuels ou des annotations. Le but est de s'élever d'un niveau d'abstraction pour ne décrire que des éléments communs à plusieurs images. Construire un vocabulaire visuel est essentiel pour représenter un ensemble d'image dans un même espace, et pour réduire les dimensions de cet espace. C'est une étape indispensable car sinon les images ne partagent aucune caractéristique. C'est une étape critique car on n'a jamais la garantie d'obtenir un vocabulaire qui ait une véritable "signification".

Notre vocabulaire visuel est construit en deux étapes. Premièrement, nous regrouperons les vecteurs de caractéristiques similaires en groupes par un algorithme d'ap-

5. Difference of Gaussian (DoG)
6. Scale Invariant Feature Transform
7. Luminance est extrait a partir de l'espace du couleur L*a*b

prentissage non-supervisé des $k - moyens$ avec k le nombre de groupe. Chaque groupe représente un terme visuel. La valeur k fixe la taille du vocabulaire visuel. Cette transformation *discrétise* la représentation quasi continue des caractéristiques en une espace discret de termes visuels. Ce vocabulaire visuel est noté $\mathcal{V}$.

La deuxième étape consiste à représenter chaque image dans ce nouvel espace des termes visuels. On construit ensuite un vecteur de terme visuel dans lequel les valeurs représentent les probabilités de distribution des termes visuels dans l'image.

Enfin, les vecteurs des termes visuels des images sont concaténé pour former une matrice de co-occurrence document/terme-visuel $M_{d,v}$. Cette matrice capture la probabilité jointe de la fréquence de co-occurrence d'une terme visuel v dans vocabulaire visuel $\mathcal{V}$ pour un document d de la base de document multimédia $\mathcal{D}$.

2.3. *Le vocabulaire textuel*

Le vocabulaire textuel est en fait l'ensemble des termes des annotations des images. Nous utilisons le modèle vectoriel sur ce vocabulaire $\mathcal{T}$ de la collection des documents D. Nous utilisons la notion de *terme* pour le texte en parallèle avec la notion *terme visuel* pour les images. Chaque document est représenté par un vecteur contenant l'histogramme de la répartition des termes dans l'annotation. Ces vecteurs sont concaténés dans la matrice de document-terme $M_{d,t}$. Chaque élément dans la matrice $M_{d,t}$ est pondéré selon le $tf \times idf$, c'est à dire la fréquence de terme (*tf*) et l'inverse de la fréquence de document (*idf*)

$$\forall d \in \mathcal{D}, \forall t \in \mathcal{T} : \begin{cases} tf_{d,t} = \frac{M_{d,t}}{\sum_{j=0}^{|\mathcal{T}|} M_{d,j}} \\ idf_{d,t} = log(\frac{|\mathcal{D}|}{tf_{d,t}}) \end{cases}$$

3. Modèle de fusion par sémantique latente

Dans cette partie, nous montrons comment les deux vocabulaires sont fusionnés à l'aide de la technique par sémantique latente.

3.1. *Analyse par sémantique latente*

La *Latent Semantic Analysis* (LSA) ou *Latent Semantic Indexing* (LSI) a été introduit initialement dans le domaine de la recherche information par Deerwester et al. [DEE 90] et par Landauer et al. [LAN 98]. Cette technique consiste à réduire la matrice d'indexation dans une nouvelle espace sensée exprimer des dimensions plus "sémantiques". Cette réduction est donc pour objectif de faire apparaître la *sémantique cachée*‘ dans les liens de co-occurrence. On parle alors de *sémantique latente*. Cette sémantique latente permet par exemple de réduire les effets de la *synonymie* et de la

polysémie. Elle est également utilisée pour indexer sans traduction, ni dictionnaire, des corpus parallèles, c'est à dire composées de documents dans différentes langues, mais sensés être des traductions les uns des autres.

Techniquement, la méthode LSA est une opération de transformation de la matrice M de co-occurrence entre les termes et les documents. Il s'agit en fait d'une *Décomposition aux valeurs singulières*[8] de la matrice M : $M_{i,j}$ décrit les occurrences du terme i dans le document j. Le but est de calculer les matrices U, Σ et V telles que :

$$M = U\Sigma V^t$$

où

$$\begin{cases} U \text{ est la matrice des vecteurs propres de } MM^t \\ V^t \text{ est matrice des vecteurs propres de } M^tM \\ \Sigma \text{ est la matrice diagonale } r \times r \text{ des valeurs singulières} \end{cases}$$

Cette transformation permet de représenter la matrice M comme un produit de deux source d'informations différentes : la matrice U relative aux documents et la seconde matrice ΣV^t relative aux termes. En utilisant les k plus grandes valeurs propres de Σ et en tronquant les matrices U et V en conséquence, on obtient une approximation de rang k de M :

$$M_k = U_k\Sigma_k V_k^t$$

où $k < r$ est la dimension de l'espace latent. Cette réduction de dimension permet de capturer l'information importante et d'éliminer l'information moins importante considérée comme du bruit produit par la redondance d'information, comme la synonymie ou la polysémie. Il faut noter que le choix du paramètre k est difficile car il doit être suffisamment grand pour ne pas perdre d'information, et suffisamment petit pour jouer son rôle de réduction de la redondance.

3.2. *Recherche de Document Multimédia (RDM)*

Chaque modalité des documents (texte et image) est traitée indépendamment (cf. figure 2). On obtient une matrice document-terme textuel $M_{d,t}$ et une matrice document-terme visuel $M_{d,v}$. La fusion de ces deux modalités est d'abord obtenue tout simplement par la concaténation des colonnes des deux matrices $M_{d,t}$ et $M_{d,v}$ en une matrice $M_{d,vt}$ car il s'agit de différentes cordonnées sur un même ensemble de documents. Cette matrice fusionnée est ensuite projetée dans un espace latent pour obtenir la matrice latente $M_{d,k}$, avec k, la nouvelle dimension réduite. Dès lors, chaque document est représenté par une ligne de la matrice latente $M_{d,k}$.

Pour une requête contenant du texte et des images, nous appliquons le même processus qu'avec les documents. Puis, ce vecteur est projeté dans l'espace réduit pour obtenir un pseudo-vecteur q_k : $q_k = q * \Sigma_k V_k^t$. Enfin, le calcul de la valeur de pertinence d'un document à la requête (Relevance Status Value ou RSV) est calculé en

8. Singular Value Decomposition (SVD)

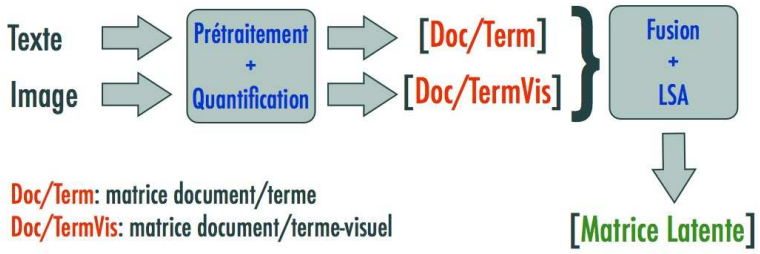

Figure 2. *Modèle de fusion par LSA pour la recherche de documents multimédias*

fonction de la similarité du vecteur requête q_k avec les lignes de la matrice latente à l'aide de la fonction $cosinus$.

3.3. *Annotation Automatique d'Images (AAI)*

L'annotation automatique des images (AAI) est assez similaire à l'indexation décrite dans la section précédente. En entrée, nous avons toujours les documents multimédias. Cependant, l'AAI consiste à trouver une corrélation entre des informations textuelles et des informations visuelles pour pouvoir attribuer automatiquement des annotation textuelles à de nouvelles images. Pour cela, il nous faut calculer la probabilité jointe entre les termes et les termes visuels.

Nous avons une approche similaire avec la méthode décrit dans [MOR 99] : la matrice de co-occurrence terme-terme visuelle $M_{t,v}$ est obtenu comme suit : (1) les images sont traitées comme en 2.1) pour obtenir un vecteur des termes visuels ; (2) chaque terme visuel est associé avec tous les termes textuels associées à l'image ; (3) les fréquences de termes sont finalement calculées pour former la matrice co-occurrence terme-terme visuelle.

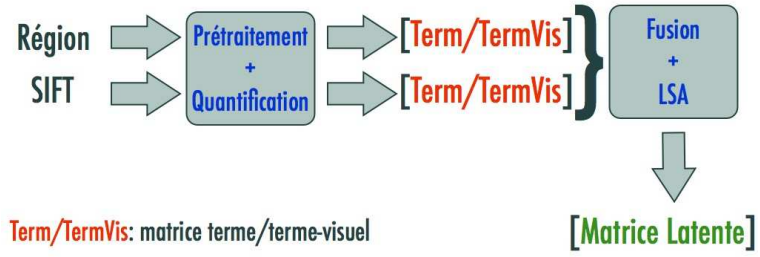

Figure 3. *Modèle de fusion par LSA pour l'annotation automatique d'images*

Cette méthode est appliquée indépendamment pour chaque type de caractéristique visuelle (i.e. région et point d'intérêt SIFT). Chaque caractéristique produit une ma-

trice terme-terme visuel particulière : $M_{t,v}^{REG}$ et $M_{t,v}^{SIFT}$. Rappelons que notre objectif est de fusionner différente caractéristiques visuelles, donc pour ce faire, les deux matrices $M_{t,v}^{REG}$ et $M_{t,v}^{SIFT}$ sont concaténées par colonnes pour former la matrice $M_{t,v}^{RS}$. C'est alors cette matrice qui est projeté dans l'espace latente par LSA pour obtenir la matrice réduite $M_{t,k}^{RS}$, où k est la dimension réduite des termes visuels.

Pour une image non-annotée, un vecteur q de fusion est obtenu en concaténant les deux vecteurs de chaque caractéristiques. On projette ensuite ce vecteur dans l'espace latent pour obtenir un pseudo-vecteur q_k de dimension réduite k. Finalement, la liste des mots-clés est calculée pour cette image, en se basant sur la fonction $cosinus$ du vecteur q_k et la matrice latente de co-occurrence de terme-terme visuel $M_{t,k}^{RS}$.

4. Évaluation expérimentale

Nous avons utilisé la collection de COREL pour une première expérimentation. Cette collection est composée de 5000 documents multimédias répartis dans 50 classes. Chaque document est associé à un ensemble de 1 à 4 termes en anglais (par exemple : "sky", "building", "sunset", "beach"). Le vocabulaire textuel de cette collection est réduit à seulement 374 termes différents. Nous avons divisé la collection en deux ensembles : l'ensemble d'entraînement constitué de 4500 documents, et l'ensemble de test constitué des autres 500 documents de la collection. Nous avons réalisé deux types de tests de notre modèle décrit dans la section 3. Notre objectif est de mesurer l'effet de l'analyse par sémantique latente sur deux tâches : RDM et AAI.

Figure 4. *Exemple de requête de la collection CLEF : "straight road in the USA".*

La seconde collection de document est la collection d'IAPR TC-12 Benchmark [GRU 06] de CLEF. Cette collection comporte 20000 images touristiques. Elle inclut des images de sports, des photographies de personnes, d'animaux, de villes. Ce sont des images prises par une agence de voyage pour ses clients. Ce sont donc des images plus "naturelles" que celle de la collection COREL. Elles sont annotées manuellement par le personnel de cette agence de voyage. L'initiative CLEF qui nous fournit cette collection, nous fournit également 60 requêtes composées de 3 images et d'une titre (voir figure 4). C'est une collection de test de RI multilingue et multimédia. En effet, les requêtes sont disponibles en plusieurs langues. Dans nos expérimentations, nous ne nous intéressons qu'à l'anglais.

4.1. *La tâche RDM*

Dans la première expérimentation avec la collection COREL, nous avons mesuré la performance du système par le "Mean Average Precision" (MAP)[9]. Le tableau 2 donne les résultats avec un modèle vectoriel standard (VSM) puis avec la sémantique latente (LSA). Les 500 images de test ont été utilisées comme requête. Un document est considéré pertinent s'il appartient à la même classe que la requête. Nous remarquons une augmentation des résultats avec la technique LSA en terme de MAP dans tous trois cas : le texte seul, l'image seule ou la fusion texte et image. Particulièrement, dans le cas de fusion les termes visuel et textuel l'usage de la LSA, améliore les résultat d'environ de 10%.

Tableau 2. *Les valeurs MAP pour différents méthodes et modalités*

Modalité	VSM	LSA	%
Image	0,1194	**0,1256**	+5,2%
Texte	0,4107	**0,4413**	+7,5%
Image + Texte	**0,4263**	**0,4694**	**+10,1%**

La figure 5 montre l'amélioration significative par la LSA en termes de courbes de précision/rappel. Nous tirons deux conclusions à partir de ces résultats. Nous remarquons tout d'abord, une grande différence de résultats entre les systèmes utilisant le texte et l'image. Du fait que les courbes $IMAGE\_TEXTE$ et $IMAGE\_TEXTE\_LSA$ sont nettement au dessus, respectivement des courbes $IMAGE$ et $IMAGE\_LSA$, nous pouvons déduire que l'apport du texte dans les systèmes fusionnant les deux modalités est très significatif. Cependant, la fusion de deux modalités peut apporter une amélioration très légère pour le texte.

Deuxièmement, nous constatons que les courbes des systèmes utilisant LSA sont au dessus de leurs homologues sans LSA. Cependant cette différence varie grandement entre les systèmes. En effet, pour ce qui est du modalité $IMAGE$ et du modalité $TEXTE$, l'amélioration observée en ajoutant LSA n'est pas très importante (5,2% pour $IMAGE$ et 7,5% pour $TEXTE$). Mais force est de constater que la plus grande différence est atteinte entre les courbes $IMAGE\_TEXTE$ et $IMAGE\_TEXTE\_LSA$, ce qui tend à prouver que LSA est d'autant plus utile que deux modalités différentes apparaissent dans l'index et sont fusionnées.

Nous avons effectué la deuxième expérience sur la deuxième collection de 20000 photos d'ImageCLEF pour confirmer l'effet du LSA sur un plus grand ensemble de données. Dans ce test, seul des termes visuels sont utilisés. Les images sont découpées en un pavage régulier de 5x5 rectangles. Puis les caractéristiques visuels suivantes sont extraites pour chaque rectangle : un histogramme de couleur de RGBL et un

9. Précision moyenne

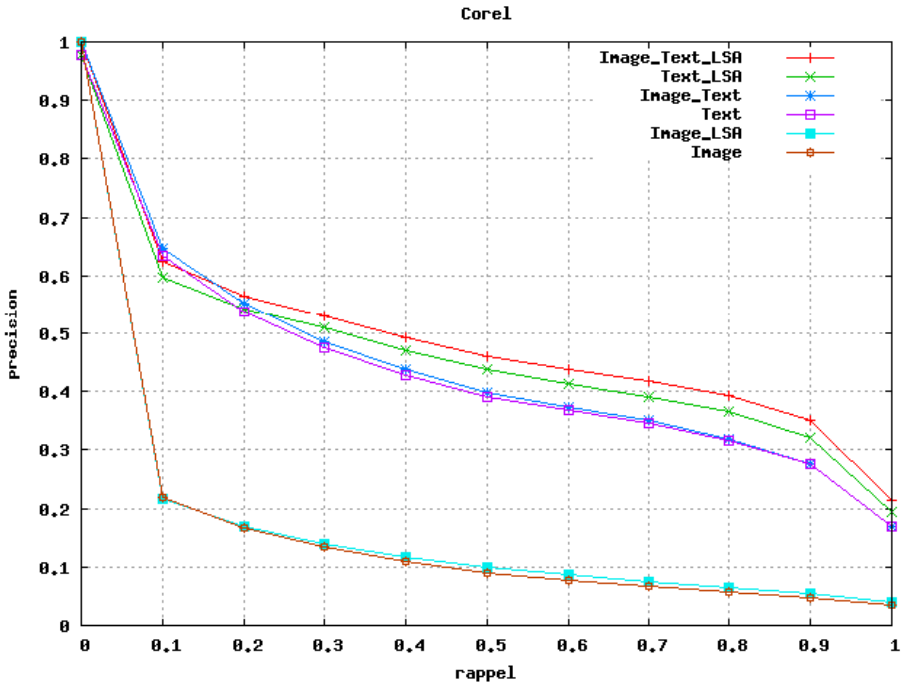

Figure 5. *Les courbes de précision/rappel pour différentes méthodes*

histogramme de bord de Canny. On applique ensuite l'algorithme de regroupement $k - moyens$ sur des vecteurs caractéristiques afin de constituer le vocabulaire visuel. Le vocabulaire est obtenu à partir de 4000 termes visuels. Le nombre de termes a été déterminé par le nombre de mots-clés dans le vocabulaire textuel des documents.

Pour comparer la méthode par sémantique latente avec la méthode vectorielle de base (VSM), nous avons fait varié la valeur de variables latentes pour trois modèles de LSA avec les trois paramètres correspondants $k1 = 100$ (LSA1), $k2 = 200$ (LSA2) et $k3 = 400$ (LSA3). Le tableau 3 montre les valeurs de MAP et la précision à 20 documents (P20) des quatre expériences.

Tableau 3. *Des valeurs MAP et P20 ont été mesurées pour différents paramètres*

	MAP	%	P20	%
VSM	0,0291		0,1417	
LSA1	0,0501	+72,1%	0,1650	+17,0%
LSA2	0,0501	+72,1%	0,1800	**+27,6%**
LSA3	0,0596	**+100,4%**	0,1717	+21,2%

Le nombre important d'images de la collection ImageCLEF 2006, rend plus difficile la tâche de recherche images basée seulement sur des caractéristiques visuelles. Cependant, nos trois expérimentations de LSA ont surpassé la méthode de base dans tous les cas (avec une amélioration de 100,4 % dans modèle LSA3 et de 72,1 % pour modèle LSA1 et LSA2). Les valeurs P20 sont également améliorées par 27,6 % avec le modèle LSA2.

Ces résultats confirment les effets positifs d'une indexation par la technique LSA sur un système de recherche d'images avec une plus grande quantité de données que les expériences précédentes. Cependant, la technique LSA a quelques limites. Par exemple, la valeur des variables latentes pour chaque système est choisi empiriquement. La technique LSA ne permet pas l'apprentissage par renforcement c.à.d. nous devons relancer entièrement le calcul chaque changement dans les paramètres.

4.2. *L'influence sur l'AAI*

Nous avons également conduit différentes expériences pour mesurer l'effet du LSA sur la tâche d'AAI. Le but est d'évaluer l'influence du LSA sur la fusion de deux modalités visuelles basées sur la segmentation en régions et la détection des points d'intérêts.

Comme le suggère [FEN 04], la performance d'un système d'AIA peut être mesuré par la *précision* et le *rappel* de chaque mot-clé. Cette méthode d'évaluation est intéressante, dans le sens où elle exprime que la tâche d'AIA est similaire à un processus de recherche d'information. Étant donné : A, le nombre d'images annotées par mot-clé w, B le nombre d'images correctement annotées par mot-clé w et C le nombre d'images dans le "ground-truth" contenant mot-clé w, la *précision* et *rappel* se calculent par la formule suivante : $R = \frac{B}{C}, P = \frac{B}{A}$.

Pour évaluer la performance du système, nous calculons la précision et rappel moyen de tous les mots dans le vocabulaire. Le but est d'obtenir une haute précision pour toutes les valeurs de rappel. Nous avons utilisé cette mesure pour comparer les 3 systèmes d'AAI qui ont été construits. Nous avons également comparé notre résultat avec la méthode classique d'AAI de traduction [DUY 02] proposé par Duygulu et al.

Quatre expérimentations AAI ont été effectuées. Elles correspondent aux différentes méthodes de construction du vocabulaire visuel et aux différentes méthodes de fusion : basé sur la région , basé sur SIFT, fusion vectorielle et fusion par LSA. Ces nouvelles expérimentations sont effectuées avec la base COREL. Nous avons construit 1000 termes visuels a partir des 4500 images d'entraînement (500 termes visuel pour les régions et 500 termes visuels pour SIFT). La matrice de co-occurrence terme/terme-visuel a une taille de 374x1000. Nous avons fixé le nombre de variables latentes à $k = 100$ ce qui réduit la matrice fusionnée à la taille de 374x100. Cette réduction permet également un gain de calcul appréciable .

Tableau 4. *Comparaison du résultat des différents systèmes d'AAI*

Model	TM	FUSION	**LSA**
Nombre de termes avec un rappel ≥ 0	49	226	**224**
Résultat sur les 49 meilleurs termes			
Rappel moyen par terme	0,34	0,45	**0,43**
Précision moyenne par terme	0,20	0,36	**0,33**
Résultat sur tous les termes			
Rappel moyen par terme	0,04	0,10	**0,09**
Précision moyenne par terme	0,06	0,08	**0,07**

Le tableau 4 présente les résultats de trois expérimentations AAI : Translation Machine (TM) [DUY 02], fusion en direct (FUSION), et fusion par sémantique latente (LSA). Nous constatons que tous les types de fusion produisent les meilleurs résultats : ils sont meilleurs que la technique de base TM. Les deux techniques de fusion donnent des valeurs moyennes de *précision/rappel* assez proches. Cela indique un effet intéressant de la sémantique latente : elle réduit la taille de la matrice tout en conservant les performances.

Tableau 5. *Les résultats d'AAI par différentes méthodes*

Image				
Human	beach people sunset water	coral ocean reefs	cars formula tracks wall	locomotive train smoke railroad
SIFT	petals swimmers leaf black pool	sphinx man girl statue woman	frost arch ice house coral	buddhist white-tailed lily deer roofs
REGION	**sunset** sea sunrise shadows tables	**reefs coral ocean** fan bridge	**formula** bengal log **tracks** head	**railroad** leaf **train** plants **locomotive**
LSA	light **sunset** sun reflection clouds	**reefs coral ocean** fish bridge	forest log cat tiger **tracks**	blooms nest marine **railroad train**
FUSION	**sunset** island palm sunrise **beach**	**coral ocean reefs** fish fan	**formula tracks** arch bridge **cars**	**locomotive train railroad** nest leaf

Le tableau 5 illustre quelques exemples d'images annotées par différents systèmes d'AAI. Les résultats d'annotation par fusion de deux modalités d'image montrent clairement l'amélioration de la qualité d'annotations de ceux qui emploient seulement une modalité d'image. Nous notons également que le modèle SIFT donne des annotations plus fréquemment incorrectes. Cela est probablement du à la complexité des descripteurs locaux pour représenter des photos.

5. Conclusion

Dans cet article, nous avons étudié l'influence de la sémantique latente pour la Recherche Document Multimédia et pour l'Annotation Automatique d'Images. Nous avons conduit des expérimentations sur des base de documents contenant un nombre significatif d'images : 5000 images dans COREL et 20000 images dans ImageCLEF 2006. C'est la première fois que cette technique est appliquée sur de larges bases et avec diffèrent types de caractéristiques visuelles. L'effet positif de l'utilisation de la technique LSA est alors clairement confirmé, avec une amélioration significative d'environs 10% de la valeur MAP lorsque l'on considère la fusion des deux modalités texte et image. Pour la tâche d'AAI, la technique de LSA améliore notablement la qualité de l'annotation en combinant différentes caractéristiques visuelles (i.e. descripteur local et caractéristique global). En complément, l'utilisation de LSA permet de réduire la complexité des calculs sur les grandes matrices tout en maintenant de bonnes performances.

Nous avons l'intention d'utiliser cette technique dans un contexte d'indexation par concepts pour une base d'images médicales [LAC 07]. Une autre piste intéressante consiste à mettre plus clairement en évidence les d'associations entre termes textuels et des termes visuels. En effet, la technique LSA est globale et ne permet pas d'analyser manuellement les associations réalisées par la réduction de dimensions.

Remerciements

Les auteurs tiennent à remercier le laboratoire IPAL pour son soutien de ce travail. Pham Trong Ton tient à remercier le support du programme Merlion de l'ambassade de France à Singapour pour son séjour doctorant en France.

6. Bibliographie

[CHE 06] CHEVALLET J.-P., MAILLOT N., LIM J.-H., « Concept Propagation Based on Visual Similarity. Application to Medical Image Annotation », *Third Asia Information Retrieval Symposium, AIRS 2006, Poster Session, Singapore*, October 2006, p. 514–521.

[COM 02] COMANICIU D., MEER P., « Mean Shift : A Robust Approach Toward Feature Space Analysis », *IEEE Trans. on PAMI*, vol. 24, n° 5, 2002, p. 603–619.

[DEE 90] DEERWESTER S., DUMAIS S., FURNAS G. W., LANDAUER T. K., HARSHMAN R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, 1990, p. 391–407.

[DUY 02] DUYGULU P., BARNARD K., DE FREITAS J., FORSYTH D., « Object Recognition as Machine Translation : Learning a Lexicon for a Fixed Image Vocabulary », *ECCV*, 2002, p. 97–112.

[FEN 04] FENG S., LAVRENKO V., MANMATHA R., « Multiple Bernoulli Relevance Models for Image and Video Annotation. », *Proc. of IEEE CVPR*, 2004.

[GRU 06] GRUBINGER M., CLOUGH P., MULLER H., DESELAERS T., « The IAPR TC-12 Benchmark : A New Evaluation Resource for Visual Information Systems. », *Proceedings of International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval*, 2006.

[INO 04] INOUE M., « On the need for annotation-based image retrieval », *Workshop on Information Retrieval in Context*, 2004, p. 44-46.

[LAC 07] LACOSTE C., LIM J., CHEVALLER J.-P., LE T., « Medical Image Retrieval based on Knowledge-Assisted Text and Image Indexing », 2007.

[LAN 98] LANDAUER T., FOLTZ P., LAHAM D., « Introduction to Latent Semantic Indexing », *Discourse Processes*, vol. 25, n° 5, 1998, p. 259–284.

[LAV 03] LAVRENKO V., MANMATHA R., JEON J., « A model for learning the semantics of pictures », *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS*, 2003.

[LI 03] LI J., WANG J. Z., « Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, n° 9, 2003, p. 1075–1088, IEEE Computer Society.

[LOW 99] LOWE D., « Object Recognition from Local Scale-Invariant Features », *Proc. of IEEE ICCV*, 1999, p. 1150–1157.

[MON 03] MONAY F., GATICA-PEREZ D., « On Image Auto-Annotation with Latent Space Models. », *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, 2003.

[MOR 99] MORI Y., TAKAHASHI H., OKA R., « Image-to-word transformation based on dividing and vector quantizing images with words », *Proc. of Intl. Workshop on Multimedia Intelligent Storage & Retrieval Mgt.*, 1999.

[QUE 05] QUELHAS P., MONAY F., ODOBEZ J.-M., GATICA-PEREZ D., TUYTELAARS T., GOOL L. V., « Modeling Scenes with Local Descriptors and Latent Aspects », *IEEE ICCV*, 2005, p. 883–890.

[SME 00] SMEULDERS A. W. M., WORRING M., SANTINI S., GUPTA A., JAIN R., « Content-Based Image Retrieval at the End of the Early Years », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, n° 12, 2000, p. 1349–1380, IEEE Computer Society.

[ZHA 02] ZHAO R., GROSKY W., « Narrowing the semantic gap - improved text-based web document retrieval using visual features », *IEEE Trans. on Multimedia*, vol. 4, n° 2, 2002, p. 189–200.