

Latent Semantic Fusion Model for Image Retrieval and Annotation

Trong-Ton Pham, Nicolas Eric Maillot,
Joo-Hwee Lim, Jean-Pierre Chevallet

Image Perception, Access & Language Lab (UMI CNRS 2955, I2R, NUS, UJF)
Institute for Informcomm Research (I²R)
21, Heng Mui Keng Terrace, Singapore, 119613.
{ttpham, nmaillot, joohwee, viscjp}@i2r.a-star.edu.sg

ABSTRACT

This paper studies the effect of Latent Semantic Analysis (LSA) on two different tasks: multimedia document retrieval (MDR) and automatic image annotation (AIA). 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.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*content analysis and indexing*

General Terms

Algorithm

Keywords

image indexing and retrieval, automatic annotation, latent semantic indexing, multimedia fusion

1. INTRODUCTION

Searching for images using their visual contents is a challenging task. Throughout nearly two decades of research on Content-Based Image Retrieval (CBIR), many different methods [14] have been proposed. But, without the explicit knowledge on what users want to search, current CBIR systems have shown limited success. Human tends to associate images with high-level concepts in everyday life. However, what the current computer vision techniques can extract

from images are mostly low-level features. The link between low-level features and high-level semantic of image content is lost. This well-known problem is called *semantic-gap* [15].

On the other hand, Annotation-Based Image Retrieval (ABIR) system [5] incorporates more efficient semantic content into both text-based queries and image captions (ie. Google Image Search, Yahoo! Image Search). Hence, the need for an Automatic Image Annotation (AIA) system integrated with current ABIR systems becomes important in the near future.

Influenced by machine learning research, there are two main approaches for the problem of image annotation and retrieval. The first one defines annotation and indexing as a supervised learning process. Training images are manually classified into a set of classes (i.e. each class corresponding to a word), or concepts (i.e. set of related keywords). A binary classifier was then trained to detect that class (concept). For a new image, the visual similarity to each class will be computed. Finally, images retrieved (annotations) will be propagated with respect to the presence or absence of the appropriating classes (concepts).

On the contrary, the second approach attempts to discover the statistical link between visual features and concepts automatically using unsupervised learning methods [9, 8, 3]. This method does not require labeled data for training the system. The idea is to introduce a set of latent variables that encode the joint distribution of words and image elements (e.g. block-based decomposition or region segmentation). Given a new image to annotate or as a query, visual features are extracted, likelihood function will return the state that maximize the joint density of semantic labels and visual elements. Images retrieved (keywords) are ranked based on the likelihood values.

The efficiency of Latent Semantic Analysis (LSA) has been proved for textual indexing [7] and visual indexing [13]. To the best of our knowledge, no work has studied its effect on a *significant* multimedia document collection for indexing and retrieval purposes. In [15], the influence of LSA has been measured on a small set of 20 documents. Recently, Monay et al. [11] has extended the experiments to a bigger collection of 8000 images and shown encouraging results of using LSA model.

The computer vision and image retrieval communities have proposed different ways for image representation: *patches* [12], *Blobword* [2] or *keypoints* [10]. In [13], an interesting

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM CIKM'07, November 6–8, 2007, Lisboa, Portugal.

Copyright 2007 ACM 978-1-59593-803-9/07/0011 ...\$5.00.

study of the influence of the latent aspects on scene representation by keypoints is proposed. Nevertheless, a lot of work remains to be done in order to study the effect of LSA on the combination of different image representations.

In this work we study the effect of LSA on the following tasks: Multimedia Document Retrieval (MDR) and Automatic Image Annotation (AIA). The contributions of this work are twofold. First, this work is an extensive study on the effect of LSA on multimedia document (i.e combination of text and image) indexing and retrieval on a significant number of documents (e.g. 20000 images come from ImageCLEF 2006 collection [4]). Second, we show that the influence of fusing several image representation methods (i.e. Blobword and local interest point) can improve the results on AIA task.

The paper is structured as follows. Section 2 details the proposed model. This section first explains how images and text are represented. Then, it is shown how these different image representations can be utilized in MDR and AIA tasks. Section 3 is dedicated to experimental results. Section 4 provides a conclusion and an overview of possible future work.

2. FUSION MODEL WITH LSA

LSA was first introduced as a text retrieval technique, motivated by problems in textual retrieval. A fundamental problem was that users wanted to retrieve documents on the basis of their conceptual meanings, and individual terms provide little reliability about the conceptual meanings of a document. This issue has two aspects: synonymy and polysemy. *Synonymy* describes the fact that different terms can be used to refer to the same concept. *Polysemy* describes the fact that the same term can refer to different concepts depending on the context of appearance of the term. LSA is said to overcome this deficiencies because of the way it associates meaning to words and groups of words according to the mutual constraints embedded in the context which they appear.

LSA relies on the following mathematical operations: a term-by-document matrix M ($M_{i,j}$ is the number of occurrences of term j in document i) of rank r is decomposed into 3 matrices by Singular Value Decomposition (SVD) as:

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

where

$$\begin{cases} U : \text{matrix of eigenvectors derived from } MM^t \\ V^t : \text{matrix of eigenvectors derived from } M^tM \\ \Sigma : r \times r \text{ diagonal matrix of singular values } \sigma \\ \sigma : \text{square roots of the eigen-values of } MM^t (M^tM) \end{cases}$$

By selecting the k largest values from matrix Σ and keeping the corresponding columns in U and V , the reduced matrix M_k is given by

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

where $k < r$ is the dimensionality of the latent space. Choosing parameter k is not a straightforward task. It should be large enough to fit the characteristics of the data. On the other hand, it must be small enough to filter out the irrelevant representation details.

2.1 Image Representation

Images are represented as bags of visual terms (*visterms*) [13] which can be regions, patches and keypoints (Fig. 1). An image is partitioned in two different ways: by patch extraction (i.e. divide into 5×5 patches) [12] and by region segmentation [2] using mean-shift algorithm [1]. Keypoints are obtained by scale-space extrema detection after filtering of the image by Difference of Gaussians operator. SIFT features [10] are then extracted from these keypoints.

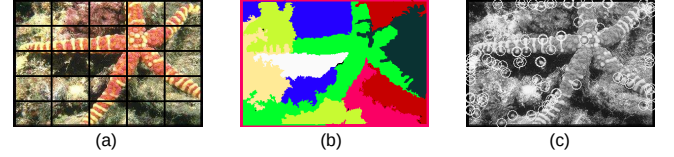


Figure 1: Different representations of an image used in our implementation: (a) grid partitioning, (b) region segmentation, and (c) local keypoints

The features extracted from regions are color histograms, Gabor coefficients, and position of the centroid. For color histogram, we augment the RGB channel with the L channel from the L^*a^*b color space. The same features are extracted from patches. The dimensions for the feature vectors are summarized in Table 1.

Table 1: Visual features used in our implementation

Feature	Quantization	Dimension
SIFT	(4×4) grids $\times$ 8 orientations	128
RGBL	16 cells $\times$ 4 channels	64
Gabor	6 orientations $\times$ 5 scales	30
Location	(x_c, y_c)	2

Bags of visterms are built as follows: (1) Unsupervised learning with k-means clustering groups similar feature vectors, extracted from regions or keypoints, to the same cluster i.e. visterm. Clustering transforms the continuous feature space into a discrete number of clusters as visual vocabulary $\mathcal{V}$. (2) Image representation consists of quantizing each image into a numerical vector representing the number of occurrences of each visterm in the image. Finally, we concatenate these vectors as rows to form a document-visterm (i.e. image-visterm) matrix $M_{d,v}$. This matrix captures the joint probabilities of the co-occurrences of visterms v in visual vocabulary $\mathcal{V}$ with each document d in a set of documents $\mathcal{D}$.

2.2 Text Processing

Based on a textual vocabulary $\mathcal{T}$, extracted from a set of documents $\mathcal{D}$, a classical vector-space model is used for text indexing and retrieval. Here, we use the notion *textterm* to refer to the keywords in $\mathcal{T}$, parallel to the use of visterm. A document-textterm matrix $M_{d,t}$ is built with each entry set to the product of the normalized versions of term frequency and inverse document frequency, $tf_{d,t} \cdot idf_{d,t}$,

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

2.3 Multimedia Document Retrieval with LSA

Each modality of a multimedia document (text and image) is processed independently (shown as two arrows in Fig. 2) to obtain the document-textterm ($M_{d,t}$) and document-visterm ($M_{d,v}$) matrices. The two modalities are fused by concatenating the columns of the two matrices as matrix $M_{d,vt}$, which is then projected into the latent space to obtain the reduced dimension $M_{d,k}$. Each multimedia document is now represented by a corresponding row in the reduced matrix. For a query with both text and image, a vector representing the number of visterns and texterns is computed. After projecting the query vector into the latent space, the *cosine* similarity is computed for each indexed document for ranking.

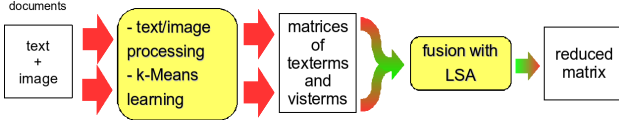


Figure 2: LSA model for multimedia document indexing

2.4 Automatic Image Annotation with LSA

Similar to multimedia document indexing and retrieval, the relation between documents and terms is expressed by the document-textterm and document-vistern matrices. However for the AIA problem, we need to describe the correlation between textual and visual information. Hence, we compute the joint probability of texterns and visterns as in [12]. The textern-vistern co-occurrence matrix $M_{t,v}$ is obtained as follows: (1) images with associated keywords are used for learning to formulate the bags of visterns as in section 2.1, (2) each vistern is associated with all words from its original image and (3) the frequencies of words of all visterns are accumulated.

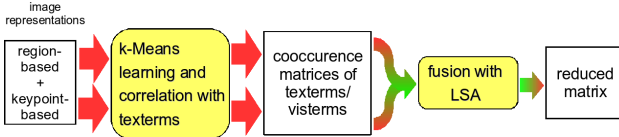


Figure 3: LSA model with multiple image representations

Two matrices representing the region-based model, $M_{t,v}^R$, and the keypoint-based model, $M_{t,v}^L$, are computed. Our goal is to fuse these two approaches using LSA over different image representations (Fig. 3). To do that, the columns of the two matrices are concatenated together to form a bigger matrix $M_{t,v}^{RL}$ and projected to the latent space using LSA to obtain a reduced matrix $M_{t,k}^{RL}$, where k is the reduced dimension from the original dimension $v^* = |v^R| + |v^L|$.

For an unannotated image, a fusion vector q is obtained by concatenating the two vectors for each image representation. The concatenated vector is then projected into the latent space to obtain a pseudo-vector, $q_k = q^t * U_k$, with the dimension reduced. After projection, the similarity between the image and words is computed using the *cosine* measure. The most probable words (e.g. top 5) will be propagated to the new image depending on the ranks of the similarities.



Figure 4: Each query topic in ImageCLEF 2006 contains 3 sample images. For instance, topic $n^{\circ}6$: *straight road in the USA*.

3. EXPERIMENTAL RESULTS

The Corel image collection was used to conduct the first experiments presented in this paper. It is composed of 5000 documents organized in 50 classes. Each document is associated with a set of keywords (between 1 and 4). The total number of keywords in the vocabulary is 374. 4500 images are used for training i.e construction of the visual vocabulary and estimation of the co-occurrence model between documents and texterns/visterns. Another 500 images are used for testing, with their associated keywords as ground truths for evaluation. Based on the model described in section 2, two types of experiments have been conducted, to measure the effects of LSA on MDR and AIA tasks respectively.

The image collection of the IAPR TC-12 Benchmark [4] consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Thanks to its well-organized and well-labeled images, this collection was used in the ImageCLEF¹ 2006 competition. Apart from the training images, query images are given for 60 different topics. Each topic contains 3 sample images (see figure 4). On this dataset, we conduct only the experiment on the effect of LSA on visual image retrieval.

3.1 Effects on Indexing and Retrieval

In the first experiment on Corel dataset, the resulting Mean Average Precision (MAP) measures are provided in Table 2. Each of the 500 images is used as a query. A retrieved document is considered relevant if it belongs to the same class as the query. Fig. 5 shows a significant improvement by LSA in terms of precision/recall curves.

Table 2: MAP for different modalities/methods

Modality	DM	LSA
Image	0.1194	0.1256 (+5.2%)
Text	0.4107	0.4413 (+7.5%)
Image + Text	0.4263	0.4694 (+10.1%)

These results lead us to the following conclusions. First, the text modality produces better results than the image modality. Nevertheless, results are slightly improved when

¹<http://ir.shef.ac.uk/imageclef/>

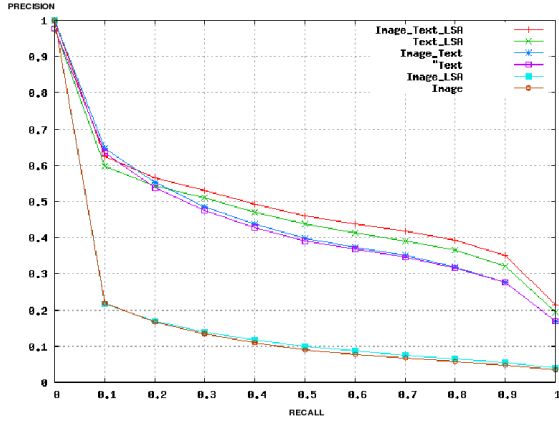


Figure 5: Precision/Recall curves for different methods

both modalities are considered. Second, LSA on both modalities lead to the best MAP and improves the MAP by 10.1% if the two modalities are considered without LSA.

Similary, we have carried out the second experiment on a 20000 tourist images dataset to confirm the effect of LSA on bigger dataset. In this test, only visual features are used. Images are divided into 5x5 non-overlapped rectangles. We extracted for each patch some visual features such as RGBL color histogram and Canny edge histogram. Patches are then clustered into 4000 clusters using K-Mean algorithm. The number of cluster has been chosen based on the number of keywords in the vocabulary. To compare with the baseline method using direct feature matching (DM), we varied the number of latent variables for three LSA models with the three corresponding parameters $k_1=100$, $k_2=200$ and $k_3=400$. Table 3 shows the MAP values and the precision values at top 20 retrieved images (P20) of four models.

Table 3: MAP and P20 values for different model

	MAP	P20
DM	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%)

As the number of images is large and the queries demand high degree semantic interpretation in ImageCLEF 2006, the task is very challenging for visual image retrieval. Overall, three LSA models outperformed the base-line result in all cases (with an improvement of 100% in model LSA3 and about 72% for both model LSA1 and LSA2). The P20 values are also improved by 27% with model LSA2.

These results confirm the good effect of LSA on image retrieval system on large scale datasets. However, LSA model is observed also some inconvenients. For example, the fact that the number of latent variables for each system is chosen empirically. In addition, the current application is not incremental (i.e. we need to retrain the system everytime if there is some changes in the internal configuration of the system).

3.2 Effects on Automatic Image Annotation

We have also conducted different experiments to measure the effect of LSA on the AIA task. The goal is to evaluate the influence of LSA on the fusion of two visual vocabularies based on regions and keypoints.

As suggested in [3], the performance of an AIA system can be measured by the *precision* and *recall* of each word. This performance evaluation method is more interesting in the sense that it expresses the AIA task as a text retrieval process. Let A be the number of images automatically annotated with a given word, B the number of images correctly annotated with that word, and C the number of images having that word in ground-truth annotation. Then *precision*(P) and *recall*(R) are computed as $R = \frac{B}{C}$, $P = \frac{B}{A}$.

Table 4: Comparison of AIA results of different methods

Model	TM [2]	DF	LSA
#words with recall ≥ 0	49	226	224
Results on 49 best words			
Mean per-word Recall	0.34	0.45	0.43
Mean per-word Precision	0.20	0.36	0.33
Results on all words			
Mean per-word Recall	0.04	0.10	0.09
Mean per-word Precision	0.06	0.08	0.07

To evaluate the system performance, the precision and recall values are averaged over the testing words. In classical information retrieval the aim is to get high precision for all values of recall. In the case of AIA, the aim is to get both high precision and high recall. We use this performance measurement to compare our results with the Translation Model (TM) [2] (see Table 4). A total of 1000 visterms has been constructed (500 region-based visterms and 500 keypoint-based visterms). A textterm/visterm co-occurrence matrix of dimension 371×1000 is then obtained. We have set the latent space dimension to $k = 100$ which leads to a matrix of dimension 371×100 . This is an important benefit from the computational point of view. Results are better than the Translation Model and are close to the results obtained without LSA.

Some examples of AIA using different methods are depicted in Table 5. The annotation results of using LSA on two image representations (last row) show clear improvement over those that use only one image representation. We also note that the keypoint-based model gives incorrect annotations more frequently probably due to the complexity of using local features in representing general photos.

4. CONCLUSION

This paper studies the effect of Latent Semantic Analysis on two tasks: *multimedia document retrieval* and *automatic image annotation*. Past studies on multimedia document retrieval have been conducted on a very small number of documents. Our study has been conducted on significantly larger number of documents i.e. 20 versus 20000. The benefit of LSA for multimedia information retrieval is clear ($> 10\%$ improvement in MAP when considering both text and image modalities).

In the task of automatic image annotation, LSA shows

Table 5: Example of AIA results by different methods

Image				
Human	beach people sunset water	coral ocean reefs	cars formula tracks wall	locomotive train smoke railroad
Keypoint	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
DF	sunset island palm sunrise beach	coral ocean reefs fish fan	formula tracks arch bridge cars	locomotive train railroad nest leaf

improvements on the combination of different image representations (i.e. local features and regions). The significant gain of using LSA in this case is the dimension reduction of the visterm/texterm vectors while maintaining similar performance.

For coping with polysemy and synonymy problems, an enhancement to our current LSA fusion model would be concept extraction for indexing [6] as we have done for medical images and text to boost the representation power of the visterms and textterms. For instance, if synonyms are mapped to the same concept, synonymy does not lead to query/document mismatching. The extraction could be driven by a thesaurus like Wordnet. We are also investigating automatic extraction of association rules between visterms and textterms.

5. ACKNOWLEDGEMENTS

We thank Clément Fleury for implementing and testing the MDR system. All systems were implemented in C++ using LTIlib² for image processing

6. REFERENCES

- [1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on PAMI*, 24(5):603–619, 2002.
- [2] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, pages 97–112, 2002.
- [3] S. Feng, V. Lavrenko, and R. Manmatha. Multiple bernoulli relevance models for image and video annotation. In *Proc. of IEEE CVPR*, 2004.
- [4] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Proceedings of International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval*, 2006.
- [5] M. Inoue. On the need for annotation-based image retrieval. *Workshop on Information Retrieval in Context*, 2004.
- [6] C. Lacoste, J. Lim, J.-P. Chevallier, and T. Le. Medical image retrieval based on knowledge-assisted text and image indexing. *IEEE Trans. on Circuit Systems and Video Technology*, 2007.
- [7] T. Landauer, P. Foltz, and D. Laham. Introduction to latent semantic indexing. *Discourse Processes*, 25(5):259–284, 1998.
- [8] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS*, 2003.
- [9] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
- [10] D. Lowe. Object recognition from local scale-invariant features. In *Proc. of IEEE ICCV*, pages 1150–1157, 1999.
- [11] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, 2003.
- [12] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proc. of Intl. Workshop on Multimedia Intelligent Storage & Retrieval Mgt.*, 1999.
- [13] P. Quelhas, D. G.-P. T. T. F. Monay, J.-M. Odobez, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *IEEE ICCV*, pages 883–890, 2005.
- [14] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.
- [15] R. Zhao and W. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *IEEE Trans. on Multimedia*, 4(2):189–200, 2002.

²<http://ltilib.sourceforge.net>