Multi Modal Semantic Indexing for Image Retrieval

Pulla Chandrika
CVIT,International Institute of Information Technology
Hyderabad-500 032, INDIA
chandrika@research.iiit.ac.in

C.V.Jawahar
CVIT,International Institute of Information Technology
Hyderabad-500 032, INDIA
jawahar@iiit.ac.in

ABSTRACT

Popular image retrieval schemes generally rely only on a single mode, (either low level visual features or embedded text) for searching in multimedia databases. Many popular image collections (eg. those emerging over Internet) have associated tags, often for human consumption. A natural extension is to combine information from multiple modes for enhancing effectiveness in retrieval. In this paper, we propose two techniques: Multi-modal Latent Semantic Indexing (MMLSI) and Multi-Modal Probabilistic Latent Semantic Analysis (MMpLSA). These methods are obtained by directly extending their traditional single mode counter parts. Both these methods incorporate visual features and tags by generating simultaneous semantic contexts. The experimental results demonstrate an improved accuracy over other single and multi-modal methods.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; I.4.10 [**Image Representation**]: Multidimensional, Statistical

General Terms

Semantic Indexing, Image retrieval, LSI, pLSA, HOSVD, SIFT

Keywords

Multi-Modal Image retrieval, Multi-Modal LSI, Multi-Modal pLSA

1. INTRODUCTION

With the popularity in large multimedia repositories over Internet increasing, need for effective access is on the rise. Most of the existing image retrieval systems use either the surrounding text or low-level features of the images to search at content-level. There is now active interest in integrating these two descriptions for building effective image retrieval systems [2, 8, 24, 36, 38]. In text based approach, images are annotated by text descriptors which are then indexed efficiently to achieve real-time retrieval [1, 26]. In this scenario, cost of annotation is very high and the whole process suffers from subjectivity of descriptors. To address this problem, content based image retrieval (CBIR) was introduced, in which images are indexed by their visual content such as color, texture, shape, spatial relationships etc. [14]. The research in this area is well established. However the effectiveness of retrieval is bottlenecked by the semantic gap [30]. That is, there is a significant gap between the high-level concepts (which human perceives) and the low-level features (which are used in describing images). Many approaches have been proposed to bridge this semantic gap between numerical image features and richness of human semantics [35].

Semantic analysis techniques are popular in multimedia processing for applications ranging from retrieval [20] to annotation [38]. Since these techniques model the concept of interest in a generic manner, they are shown to be superior to the direct feature based methods. They are very effective, when the concept of interest is complex and the number of examples is limited. Latent Semantic Indexing (LSI), probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Analysis (LDA) are the popular techniques in this direction. The basic mathematical models behind these techniques are borrowed from the the text modelling and retrieval literature [3, 9, 33]. Later they were effectively extended for vision tasks [15, 22, 29]. With the introduction of bag of words (BoW) methods in computer vision, semantic analysis schemes became popular for tasks like scene classification and segmentation [4, 37]. Visual bag of words approach represents the image as a histogram of visual words. Matching problem is then modelled as the estimation of similarities between given histograms (or probability distributions). With this modelling, it became possible to explain the image in terms of a predefined vocabulary.

The essence of semantic analysis is in decomposing the original signal/representation according to a generative process. The parameters associated with the generative process is learned from the examples. This is often achieved by a factorization scheme [33] or an Expectation Maximization (EM) based component extraction [9]. This learning process typically provides a new feature representation, which

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.
CIVR '10, July 5-7, Xi'an China
Copyright ©2010 ACM 978-1-4503-0117-6/10/07 ...$10.00.

is data dependent. Thus, it is also viewed as dimensionality reduction. Semantic indexing schemes are applied for effective search in text as well as image databases [9, 25]. However, many of the emerging databases are multimodal in nature. For example, the image collections over Internet can be effectively searched with a combination of textual and image clues. Multimodal techniques have shown prospects in many tasks including image retrieval, video search and summarization [12, 18, 38]. Zhang *et al.* [38] propose a probabilistic semantic model which generates an offline image to concept word model. An online image to text and text to image retrieval is performed in a Bayesian framework on this model. Guo *et al.* [8] proposed a multi model web image retrieval techniques based on multi-graph enabled active learning. Here, three graphs are constructed on images content features, textual annotation and hyper links respectively.

In this paper, we propose two techniques, Multi-modal Probabilistic Latent Semantic Analysis (pLSA) and Multimodal Latent Semantic Indexing (LSI). These methods incorporate both visual features and tags by generating semantic contexts. In the next sections,

- LSI is extended to Multi-modal LSI, with a tensorial representation and Higher Order SVD.
- pLSA is extended to Multi-modal pLSA by combining multiple modes into a single context, and then using EM algorithm to fit the model parameters.
- Superiority of the proposed methods is demonstrated over standard data sets. We compare our results with other methods.

1.1 Background

Semantic Indexing techniques like Latent Semantic Indexing [33], Probabilistic Latent Semantic Analysis(pLSA) [9] and Latent Dirichlet Allocation(LDA) [3] were used to improve the retrieval performance by reducing or bridging the semantic gap. These are unsupervised methods where a document is viewed as collection of words. A latent concept or hidden topic is introduced between words and documents. A generative model is first learnt, and the learnt model is then used for mapping the problem from an input space to a novel feature space. It is believed that this new representation is closer to the semantic description. In this paper, we limit our attention to LSI and pLSA. We now briefly summarize them.

Latent Semantic Analysis(LSA).

LSA was first proposed by the text retrieval community for textual indexing [33]. The basic idea is to retrieve documents based on their conceptual meaning using a term-documents matrix N. The elements of the matrix $n(d_i, w_j)$ specifies the number of times the word w_j occurred in a document d_i. Because of the semantic relationship in documents, it is argued that the term document matrix N is sparse and rank deficient, say of rank r. This term-document matrix is then decomposed into three matrices by Singular Value Decomposition (SVD). The k top largest eigenvalue values form the decomposed matrices are selected to form a reduced matrix N_k,where $k < r$ is the dimensionality of the latent space. Original data is then mapped to this reduced dimension with a linear transformation. Later Quelhas *et al.* [23] demonstrated the efficiency of LSA for visual indexing. Here the vocabulary $W = \{w_1, \ldots, w_{N_v}\}$ is formed by visual words obtained from the features extracted of images. Refer [33, 23] for detailed explanation on how LSI is extended to images.

Probabilistic Latent Semantic Analysis (pLSA).

The pLSA was originally proposed by T.Hofmann in the context of text document retrieval [9], where each document is represented as a bag-of-words representation. It has also been applied to various computer vision problems such as classification [4], images retrieval [25], where each image is considered as a single visual document and features extracted from images form visual words. The key concept of the pLSA model is to map high dimensional word distribution vector of a document to a lower dimensional *topic vector* or *aspect vector* z_k. Thus, it introduces an unobservable latent topic between the documents and the words. Each document consists of mixture of multiple topics and thus the occurrences of words is a result of the topic mixture. One of the aspect of this model is that word occurrences are conditionally independent from the document given the unobservable aspect. Thus

$$P(d_i, w_j) = P(d_i) \sum_k P(z_k|d_i) P(w_j|z_k). \qquad (1)$$

The unobservable probability distribution $P(z_k|d_i)$ and $P(w_j|z_k)$ are learned from the data using the Expectation - Maximization Algorithm (EM-Algorithm)[5]. EM algorithm is a standard iterative technique for maximum likelihood estimation, in latent variable models, where each iteration is composed of two steps (i) an Expectation (E) step where, based on the current estimates of the parameters, posterior probabilities are computed for the latent variables z_k , (ii) a Maximization (M) step, where parameters are updated for given posterior probabilities computed in the previous E step. It increases the likelihood in every step and converges to a maximum of the likelihood

Multimodal Methods.

Multimodal methods are getting popularity in image and video analysis in recent years [24, 32]. This can be partly attributed to the popularity of large multimedia repositories over web, and the user interfaces which use primarily text to search and access them. Manual annotation is getting replaced by autoannotation [16] with promising results. Auto-annotation systems assign similar keywords to similar images, which is a natural alternative to expensive and labor intensive manual annotation. Both the generative model and the discriminant methods of machine learning have been applied to learn the correlation between image features and textual words from the example of annotated images and then applied to predict the words for unseen images. Guo *et al.* [8] introduce a max margin framework on image annotation and retrieval as a structured prediction model where input and output are structures. Here, the image retrieval problem is formulated as quadratic programming (QP) problem. By solving this QP problem the dependency information between different modalities can be learned. Scenique [2] is a multimodal image retrieval system which provides integrated query facility and is based on the multi-structure framework which consists of set of objects together with schema that specifies the classification of objects according to multiple distinct criteria. The tags are organized as dimensions which take the form of tag trees.

When content based and tag based queries are given, the system return the images in intersection of content based retrieval and tag based retrieval first, followed by tag based results only, finally by image based results only.

The organization of the rest of the paper is as follows. Section 2 presents Multi Modal Latent Semantic Indexing. Section 3 presents proposed Multi Modal probabilistic Latent Semantic Analysis. Section 4 presents the retrieval schema based on the two methods proposed. Section 5 presents the comparative study of our method with current state of the art methods and presents some analysis on obtained results. Finally in section 6 we conclude.

2. MULTI MODAL LATENT SEMANTIC INDEXING

The term-document matrix is a high dimensional representation of the image in which each image is represented as frequency of the visual words. In retrieval domain most of the systems are based on direct matching of the visual words. However generally, different visual words are used to describe same concepts or different concepts are described using similar visual words because of which direct matching of visual words may not lead to efficient retrieval systems. LSI tries to search relevant documents by mapping high dimensional vector to a low dimensional latent semantic space. Thus removing the noise found in images, such that two documents that have same semantics will be located close to one another in a multi-dimensional space. Most of the current image representations either rely solely on visual features or on surrounding text.

Matrix decomposition techniques like singular value decomposition(SVD), Principal component analysis(PCA) etc are useful for dimensionality reduction, mining, information retrieval and feature selection. But these are limited to two orders only. Generally most of the data have a multidimensional structure and it is some what unnatural to organize them as matrices or vectors. For example a video is a collection of images and audio over a time stamp. Thus in many cases it is beneficial to use the available data without destroying its inherent multidimensional structure. Our tensor based model capture information for more than two orders where tensor is multidimensional or multimode arrays.

In [20], author shows the effect of LSA on Multimedia document indexing and retrieval by combining both text and image. Here, they concatenate the columns of the two matrices $N_{M\times N_t}$ and $N_{M\times N_v}$ (M number of images , N_t number textwords and N_v number of visual words in the database) into a single term document matrix and then decompose into reduced dimension to form a latent space. But this does not lead to desired improvement in retrieval results because the visual words have a much larger frequency as compared to text words. The difference in the dictionary size for the two is large as well. To overcome the above disadvantages, we propose $MMLSI$, where the data is represented by a 3-order tensor in which the first dimension is images, second is visual word and the third is the text words. Three-mode analysis using Higher Order Singular Value Decomposition (HOSVD)[11] is performed on the 3-order tensor which captures the latent semantics between multiple objects like images, low-level features and surrounding text. HOSVD technique can find some underlying and latent structure of images and is easy to implement. It helps to find correlated dimensions within the same mode and across different modes.

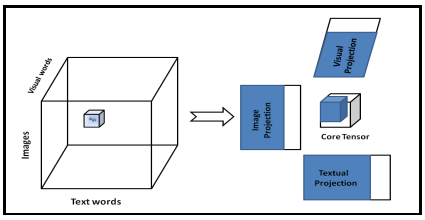

Figure 1: The figure shows visual word - text word - document tensor and its decomposition

Tensor methods have been used for a long time in chemometrics and psychometrics[31]. Recently HOSVD has been applied to face recognition [34], Handwritten digit classification[27] and data mining [10].

2.1 MMLSI

A tensor is a higher order generalization of a vector(first order tensor) and a matrix (second order tensor), also known as n-way array or multidimensional matrices or n-mode matrix. A tensor $\mathcal{A}$ can be represented as

$$\mathcal{A} \in R^{I_1\times I_2\cdots\times I_N} \tag{2}$$

The mode-d metricizing or matrix unfolding of an N^{th} order tensor $\mathcal{A} \in R^{I_1\times\cdots\times I_N}$ are vectors in R^N_d obtained by keeping index d fixed and varying the other indices. Therefore, the mode-d matricizing $A_{(d)}$ is in $R^{(\Pi_{i\neq d}N_i)\times N_d}$. See [11] for details on matrix unfoldings of a tensor.

Higher Order SVD(HOSVD) is an extension of SVD and represented as follows

$$\mathcal{A} = \mathcal{Z} \times_1 U_1 \times_2 U_2 \cdots \times_N U_N \tag{3}$$

where $U_1, U_2, \ldots, U_N$ are orthogonal matrices that contain the orthonormal vectors spanning the column space of the matrix unfolding $A_{(i)}$ with $i = 1, 2, \ldots, N$. $\mathcal{Z}$ is a *core* tensor, analogous to the diagonal singular value matrix in conventional SVD as shown in Figure 1 HOSVD is computed by the following two steps.

1. For $i = 1, 2, \ldots N$, compute the unfolding matrix $\mathbf{A}_{(i)}$ from $\mathcal{A}$ and compute its standard SVD: $\mathbf{A}_{(i)} = \mathbf{USV^H}$; the orthogonal matrix $\mathbf{U}^{(i)}$ is defined as $\mathbf{U}^{(i)} = \mathbf{U}$, i.e., as the left matrix of SVD on $\mathbf{A}_{(i)}$.

2. Compute the core tensor using the inversion formula

$$\mathcal{Z} = \mathcal{A} \times_1 \mathbf{U}^{(1)^H} \times_2 \mathbf{U}^{(2)^H} \cdots \times_p \mathbf{U}^{(p)^H} \tag{4}$$

where the symbol H denote the Hermitian matrix transpose operator.

As we are considering two modes, first we construct a tensor $\mathcal{A} \in R^{I_1 X I_2 X I_3}$ where, I_1 is the number of the images in the dataset, I_2 is the visual vocabulary size and I_3 is the text vocabulary size. Whereas, a_{ijk} is defined as number of occurrences of visual word v_j and text word t_k in a document d_i. Once the tensor is generated we decompose it by using HOSVD is shown in Figure 1 to obtain

$$\mathcal{A} = \mathcal{Z} \times_1 U_{images} \times_2 U_{visualwords} \times_3 U_{textwords}.$$

Here the, the matrices U_{images}, $U_{visualwords}$ and $U_{textwords}$ define the space of the image parameters, visual parameters and textual parameters respectively. An approximate tensor is constructed $\tilde{\mathcal{A}}$ by selecting the top k columns from the decomposed matrices. This in effect maps the data into a semantic space, which is derived from the multiple data modes. The semantic space has a lower dimension than the dictionary space. Hence in effect mapping the data into a lower dimensional space.

3. SEMANTIC INDEXING BY MULTI-MODAL PLSA

Although LSA has been successfully applied for semantic analysis for various applications like Information retrieval, image annotation and object categorizing. It has a number of disadvantages mainly due to its unsatisfactory statistical foundation. Where as, pLSA is a generative model of the data with strong statistical foundation, as it is based on the likelihood principle. It has found successful applications in single mode data such as text, image tags and visual words. In [25], author shows the dimensionality reduction due to the aspect model of pLSA which improves the performance on similarity task for a large data bases.

In a recent work [24], pLSA has been extended to multimodal data, using visual words and image tags. Here they present a probabilistic semantic model to connect image tags and visual words via a hidden layer which determines the semantic concept between the two modes. First pLSA is applied to each mode separately, and then the derived topic vectors of each mode are concatenated. pLSA is applied on top of the derived vectors to learn the final document concept relation. This is equivalent to forming an alternative dictionary of concepts, one for each mode, and merging them on which pLSA is performed. An improvement in performance is expected over naive merging of dictionaries, as the effect of difference in distribution patterns of each mode is normalized in this method. But it has an intrinsic problem of having to merge dictionaries of the different modes. This method does not place importance to interactions between the different modes. We argue that such interactions have the ability to find useful information in the dataset.

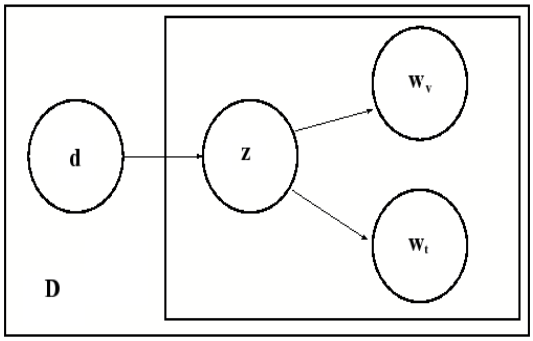

Figure 2: Graphical representation of Multi Modal pLSA

We propose a system to capture the patterns between images, text words and visual words by using EM algorithm to determine the hidden layers connecting them. An unobservable latent variable $z \in Z = z_1, \ldots, z_k$ is associated with each occurrence of the text word $w^t \in W = w_j^t, \ldots, w_{N_t}^t$ and visual word $w^v \in W = w_l^v, \ldots, w_{N_v}^t$ in a document $d \in D = d_i, \ldots, d_M$. To simplify the model, we assume that the pair of random variables (w_j^v, w_j^t) are conditionally independent given the respective image or document d_i. Thus

$$P(w_l^v|w_j^t, d_i) = P(w_l^v|d_i) \tag{5}$$

Now consider a joint probability model for text words, images or documents and visual words as

$$P(w_j^t, d_i, w_l^v) = P(w_j^t)P(w_j^t|d_i)P(w_l^v|w_j^t, d_i) \tag{6}$$

By substituting equation (5), equation (6) can be reduced to

$$P(w_j^t, d_i, w_l^v) = P(d_i)P(w_j^t|d_i)P(w_l^v|d_i) \tag{7}$$

Where, $P(w_j^t|d_i)$ probability of occurrence of text word w_j^t given a document d_i, similarly $P(w_l^v|d_i)$ probability of occurrence of text word w_l^v given a document d_i. Generally, documents consist of mixture of multiple topics and occurrences of words (i.e., visual words and text words) is a result of topic mixture. The generative model is expressed in terms of the following features:

1. pick a latent class z_k with probability $P(z_k|d_i)$.
2. generate a text word w_j^t with probability $P(w_j^t|z_k)$.
3. generate a visual word w_l^v with probability $P(w_l^v|z_k)$

The joint probabilistic model for the above generative model is given by the following:
$P(w_j^t, d_i, w_l^v)$

$$= P(d_i)\sum_k P(w_j^t|z_k)P(z_k|d_i)P(w_l^v|z_k)P(z_k|d_i) \tag{8}$$

$$= \frac{P(d_i)^2 \sum_k P(w_j^t|z_k)P(w_l^v|z_k)P(z_k|d_i)^2}{P(z_k)} \tag{9}$$

The Figure 2 shows the pictorial representation of the model. Here the a combination of text words and visual words is used to represent the image upon which higher level aspects are learned.

By following the Maximum likelihood principle we can determine $P(z_k|d_i)$, $P(w_j^t|z_k)$ and $P(w_j^v|z_k)$ by maximizing the log-likelihood function.

$$L = \Pi_{i=1}^{M}\Pi_{j=1}^{N^t}\Pi_{l=1}^{N^v}[P(w_j^t, d_i, w_l^v)^{n(w_j^t, d_i, w_l^v)}] \tag{10}$$

Taking the log to determine the log-likelihood L of the database

$$L = \sum_{i=1}^{M}\sum_{j=1}^{N^t}\sum_{l=1}^{N^v}[n(w_j^t, d_i, w_l^v)P(w_j^t, d_i, w_l^v)] \tag{11}$$

By substituting the equation (9) in equation (11) we learn the unobservable probability distribution $P(z_k|d_i)$, $P(w_j^t|z_k)$ and $P(w_j^v|z_k)$ from the data using the Expectation-Maximization Algorithm (EM-Algorithm):[5]

E-Step:

$$P(z_k|d_i, w_j^t) = \frac{P(w_j^t|z_k)P(z_k|d_i)}{\sum_{n=1}^{k} P(w_j^t|z_n)P(z_n|d_i)} \tag{12}$$

$$P(z_k|d_i.w_l^v) = \frac{P(w_l^v|z_k)P(z_k|d_i)}{\sum_{n=1}^{k} P(w_l^v|z_n)P(z_n|d_i)} \tag{13}$$

M-Step:

$$P(w_j^t|z_k) = \frac{\sum_{i=1}^{M} n(d_i, w_j^t) P(z_k|d_i, w_j^t)}{\sum_{j=1}^{N}\sum_{i=1}^{M} n(d_i, w_j^t) P(z_k|d_i, w_j^t)} \quad (14)$$

$$P(w_l^v|z_k) = \frac{\sum_{i=1}^{M} n(d_i, w_l^v) P(z_k|d_i, w_l^v)}{\sum_{l=1}^{L}\sum_{i=1}^{M} n(d_i, w_l^v) P(z_k|d_i, w_l^v)} \quad (15)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N}\sum_{l=1}^{L} n(d_i, w_j^t, w_l^v) P(z_k|d_i, w_j^t) P(z_k|d_i, w_l^v)}{n(d_i)} \quad (16)$$

The learning process is iterating the E-Step and M-Step alternatively until some convergence condition (such as Log likelihood) is satisfied. Typically, 100-150 iterations are needed before converging. Thus finally images are mapped to a lower dimensional latent vector derived from both text words and visual words. In the next section we discuss how the proposed indexing methods can be used for multi-modal image retrieval.

4. INDEXING AND RETRIEVAL

As mentioned earlier many current retrieval system depends on either text or visual features. But in many cases information available is richer and is available as a combination of different modes. For example any web page contains text, imagery and other forms of information. The research in these modalities is well established like [17] builds a system using visual words, where commercials systems like flickr user text words. The retrieval effectiveness has a bottleneck of semantic gap. In recent years, research has been done to address semantic gap problem, but these methods fail to relate an image to an abstract concept. Thus, an image retrieval system which focuses on exploiting the synergy between different modes helps in improving the retrieval efficiency.

4.1 Feature Extraction

Visual Vocabulary.

For a given image, first interest points are detected from which feature vectors are extracted. Once the features were extracted the cumulative feature space was vector quantized into clusters. These clusters form the visual words and each image is represented as a histogram of visual words.

Textual Vocabulary.

For the textual representation of each image, the keywords were extracted from the corresponding annotated text by removing stop words and stemming the remaining words. Thus for each image the key text words were found and the dataset is represented as term-document matrix. Thus, the visual words and key words forms the two modes of the documents.

4.2 Image Retrieval Framework

For a tensor based image retrieval, a multi modal framework is used to combine multiple modes to generate an image retrieval system as shown in section 2, Here, first we need to construct a tensor $\mathcal{A}$ from the dataset. Once the feature extraction is done, image are represented as histogram of visual words and histogram of keywords. A Tensor $\mathcal{A}$ is

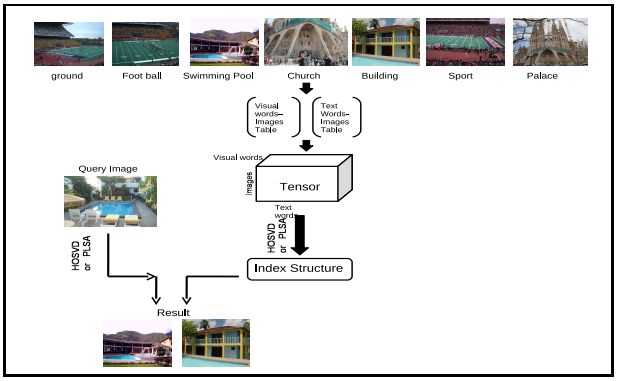

Figure 3: Over view of the Process

constructed by the following equation

$$\mathcal{A}(i,j,l) = n(d_i, w_j^t) \cdot (1-\alpha) + n(d_i, w_l^v) \cdot (\alpha).$$

Where, $n(d_i, w_j^t)$ specifies the number of time the text word w_j^t occurred in a document d_i and $n(d_i, w_l^v)$ specifies the number of times the visual word w_l^v occurred in a document d_i. This is based on the amount of information each mode has. We choose α such that the resulting matrix has a distribution which balances the effect of the multiple modes on the semantic generation. An efficient process to find an optimal α is beyond the scope of the current discussion. Then tensor $\mathcal{A}$ is decomposed using HOSVD as explained in section 2. From resulting decomposition select the top k columns to form a reduced dimensional space. The reconstructed tensors is denoted by

$$\tilde{\mathcal{A}} = \mathcal{Z} \times_1 \tilde{U}_{images} \times_2 \tilde{U}_{visualwords} \times_3 \tilde{U}_{textwords}$$

The database image and the queries are mapped on to the 2 base $\tilde{U}_{visualwords}$ and $\tilde{U}_{textwords}$. And a Euclidean distance between them is calculated to rank the relevance of the images. see Algorithm 1

Algorithm 1 Multi modal LSI

1: Construct tensor $\mathcal{A} \in R^{I_1 X I_2 X I_3}$data. Where I_1,I_2,I_3 are the numbers of image, visual words and text words respectively. Now each tensor element measures the frequency count of visual word, text word in an image.
2: Decompose the matrix using HOSVD and select the first k eigen values.
$$\mathcal{A} = \mathcal{Z} \times_1 U_{images} \times_2 U_{visualwords} \times_3 U_{textwords}$$
3: Project each image on the 2 bases $U_{visualwords}$ and $U_{textwords}$:
$$\mathcal{A}_d = U_{visualwords}^T \times \mathcal{A}^{I_1} \times U_{textwords}$$
4: Project query image on the 2 bases, derived in step 2 above, using the following:
$$\mathcal{A}_d = U_{visualwords}^T \times \mathcal{A}_{query} \times U_{textwords}$$
5: Calculate the Euclidean distance norm D between the projected image and the query.

Now, we explain the basic approach to extended pLSA for multi-modal data, using visual words and image tags. This is done by concatenating the term document matrix for image tags N_{MXN_t} and visual words N_{MXN_v} into $N_{MX(N_t+N_v)}$ and then applying standard pLSA [24]. But this does not

show any improvement in the quality of retrieval for average case scenario. The performance invariance is caused because the visual words have a much larger frequency as compared to text words and the difference in the dictionary size for the two is large. Another approach is to apply pLSA on term document matrix for image tags N_{MXN_t} and visual words N_{MXN_v} separately and then the results are combined set operations like union or intersection. The problem to determine the weights of the text and visual words is not trivial.

For image retrieval system based on Multi modal pLSA, the topic specific distributions $P(w_j^t|z_k)$ and $P(w_l^v|z_k)$ are learnt from the set of training images according to the method explained in section 3. Each training image is then represented by a Z-vector $P(z_k|d_{train})$,where Z is the number of topics learnt. Using the same approach, given a new test image d_{test} we estimate the aspect probabilities $P(z_k|d_{test})$. The probabilities $P(w_j^t|z_k)$ and $P(w_l^v|z_k)$ learned from train set are kept constant. The similarity between the test and training images is calculated using the cosine metric between the two aspect vectors $a = (P(z_k|d_{train}))$ and $b = (P(z_k|d_{test}))$. (see Algorithm 2)

Algorithm 2 Multi modal pLSA

1: • **Training Phase:**
2: Randomize and normalize $P(w_j^t|z_k)$, $P(z_k|d_i)$, and $P(w_l^v|z_k)$ to ensure the sum of all probabilities equal to one.
3: **while** not convergence **do**
4: **E-step**: Compute the posterior probabilities $P(z_k|d_i, w_j^t)$ and $P(z_k|d_i, w_l^v)$.
5: **M-step:** Parameters $P(w_j^t|z_k)$, $P(z_k|d_i)$, and $P(w_l^v|z_k)$ are updated from the posterior probabilities computed in E-step.
6: **end while**
7: • **Testing Phase:**
8: The E-step and M-step are applied on the testing data by keeping the probabilities $P(w_j^t|z_k)$ and $P(w_l^v|z_k)$ learnt from the training constant.
9: Calculate the cosine metric between the probabilities learnt from training and testing.

5. RESULTS AND DISCUSSIONS

In this section, we present the various experimental results for the proposed *MultimodeLSI* and *MultimodepLSA* on the datasets described below.

5.1 Data Sets

The following datasets are used for the evaluation of the methods proposed.

University of Washington(UW) Dataset: This dataset is used in [13] and consists of 1109 images with a ground truth of manually annotated key words. For evaluation the retrieved image is considered relevent if it belongs to the same class as the query image.

Multi-label Image Dataset: This dataset is used in [28] and consists of 139 urban scene images and four overlapping labels: *Buildings*, *Flora*, *People* and *Sky*. Each image has a minimum of two tags and each label is present in at least 60. For visual evaluation we manually created a ground truth data for 50 images.

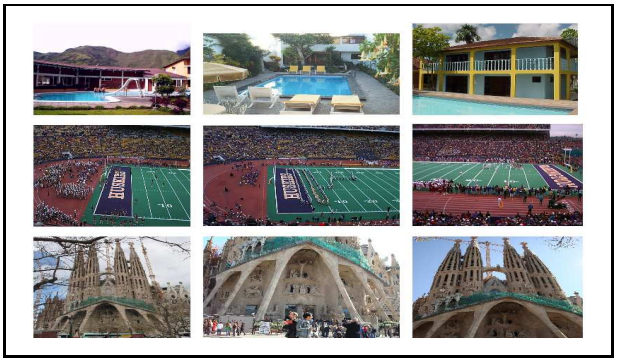

Figure 4: The first image of each row is the query, other two are the retrieved results. Each row corresponds to the IAPR, UW and Multi-label datasets respectively

IAPR TC12 Dataset: This data set consists of 20,000 images of natural scenes that include different sports and actions, photographs of people, animals, cites, landscapes and many other aspects of contemporary life. Here the images are accompanied with description in several languages and typically used for cross-language retrieval[7], we have concentrated on English captions and extracted keywords using natural language processing techniques. The vocabulary size is 291 and 17,825 images were used for training, and 1,980 for testing.

Corel Dataset: This dataset is used in [21] which consists of 5000 images out of which 4500 images are used for training and 500 image for testing. The dictionary contains around 260 uniques words. The retrieved image is considered relevent if it belongs to the same class as the query image.

5.2 Experimental Results

Initially all the images from the datasets were down sampled to reduce number of interest points, after which feature detection and SIFT feature extraction [6] is applied. For corel dataset we calculated dense sift. Now the features are vector quantized using k-means. For our experiments we created a visual vocabulary size of 500 for all the datasets, except for IAPR for which the vocabulary size is 1000. For benchmarking, we compared our method against the following classes of modes:

- **Single mode:** This refers to methods that consider only a single mode throughout the process[19, 25]. For example text only and visual words only methods lie in this category
- **Pseudo single mode:** This category of applications use single mode methods, but can use data from multiple modes. One of the methods to do so is to merge the dictionaries[21, 24]. Hence in effect considering that all the modes present in the dataset as a single mode. This merged mode is then processed by single mode methods. This is a naive way of managing multimode data. The disadvantages include shadowing of one mode by another by factors that include dictionary size, distribution etc. As these factors are crucial in the performance of single mode methods, very little

Table 1: Comparing Multi Modal LSI with different forms of LSI for all the datasets in mAP.

	visual-based	tag-based	Pseudo single mode	MMLSI
UW	0.46	0.55	0.55	0.63
Multilabel	0.33	0.42	0.39	0.49
IAPR	0.42	0.46	0.43	0.55
Corel	0.25	0.46	0.47	0.53

Table 2: Comparing Multi Modal PLSA with different forms of PLSA for all the datasets in mAP.

	visual-based	tag-based	Pseudo single mode	mm-pLSA	our MM-pLSA
UW	0.60	0.57	0.59	0.68	0.70
Multilabel	0.36	0.41	0.36	0.50	0.51
IAPR	0.43	0.47	0.44	0.56	0.59
Corel	0.33	0.47	0.48	0.59	0.59

advantage can be gained out of such a method.

- **Explicit dual mode:** These methods are designed so as to appreciate the diversity in the semantics of information represented by each mode. For example, one mode can have a small dictionary, but the distribution is such that the semantics can be easily found, another might have a much larger dictionary, but the average vocabulary per document is small. One such method present in literature is that of multi-modal multi-layer pLSA [24].

In the current context, visual words and text words are the two modes we have focused upon. For single mode methods, either of text or visual words is used. For Pseudo dual mode methods, the dictionaries are concatenated. The resulting dictionary is then used. For example, for the IAPR dataset, the visual dictionary is of size 500, and the text dictionary is of size 291, hence the resulting dictionary is of size 791, with the first 500 representing the visual words.

As discussed in the previous sections, LSI and pLSA based methods are compared in different modes. Multimodal methods are present for pLSA, hence they have also been tested. For all our experiments the number of concepts is determined by the concepts present in the respective databases which is known. We use mean Average precision (mAP) for comparison. The results of the experiments are as shown below:

LSI and variants Compared to variants of LSI, our method performs better see Table 1. It is to be noted that a better tag base has a stronger impact on accuracy of results as compared to a better visual word. This can possibly be because most key text words are found only in a very few documents, and are related to each other very strongly. Also, concatenation of the two together did not provide any appreciable performance improvements, in some cases accuracy reduced below that of tag based LSI. The values so derived are heavily biased towards the results obtained from the tags alone. Thus proving our proposition. The results obtained by our method are stronger than the other results, but on the contrary the time and space consumption for our method is much larger than the others.

PLSA and variants A similar direct comparison shows us that other than the Corel data set. The results of concatenated PLSA are dominated by the results of visual word based PLSA. Similar to the LSI models here we construct a pLSA model solely based on visual features or tags and a concatenated plSA model. Then we implemented a Fast Initialization variant of multi modal multi layer pLSA(mm-pLSA) proposed in [24]. The Table 2 shows the comparison of these methods with the proposed Multi model pLSA. Our method outperforms current single mode and multi-mode methods in performance.

From the two Tables 1 and 2 we can see that the performance of the probabilistic methods is better than the Latent semantic analysis. It can also be seen that methods that efficiently make use of multiple modes of information are able to generate better semantics. An obvious problem with such methods is the time taken to update the model given a dynamic database. Hence the focus can be on efficient methods to manage dynamic multimodal data. Thus methods that generate just in time results on a dynamic database are required.

6. CONCLUSIONS

In this paper, a direct extention to the traditional single mode semantic system to a multimodal semantic system has been proposed. Our proposed Multi modal Latent Semantic Indexing and Multi modal Probabilistic Latent Semantic Indexing systems are shown to be outperforming the state of the art. We validate our method on a number of data sets. Like pLSA and LSI, our multimodal methods are also memory and computation intensive. We are presently working on developing just in time semantic indexing for fast and effective retrieval.

7. REFERENCES

[1] www.flickr.com.

[2] I. Bartolini and P. Ciaccia. Scenique: a multimodal image retrieval interface. In *AVI '08: Proceedings of the working conference on Advanced visual interfaces*, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV '06: European Conference on Computer Vision*, 2006.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society*, 1977.

[6] D.Lowe. Distinctive image feature scale-invariant keypoints. In *IJCV:Int. J. Comput. Vision*, 2004.

[7] M. Grubinger. Analysis and evaluation of visual information systems performance. In *PhD thesis, Victoria University Melbourne, Australia*, 2007.

[8] Z. Guo, Z. Zhang, E. P. Xing, and C. Faloutsos. A max margin framework on image annotation and multimodal image retrieval. In *ICME*, 2007.

[9] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. In *Machine Learning,42,Numbers 1-2:177âĂŞ196*, 2001.

[10] T. G. Kolda and J. Sun. Scalable tensor decompositions for multi-aspect data mining. In *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008.

[11] J. V. L. De Lathauwer, B. DeMoor. A multilinear singular value decomposition. In *SIAM J. Matrix Anal. Appl*, 2000.

[12] D.-D. Le, F. Yamagishi, and S. Satoh. Video search by multi-modal and clustering analysis. In *CIVR*, 2007.

[13] C. W. Lei Zhang, Hong-Jiang Zhang. Scalable markov model-based image annotation. In *IJCV*, 2004.

[14] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2006.

[15] L.-J. Li, G. Wang, and null Li Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *CVPR*, 0:1–8, 2007.

[16] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, 2008.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[18] J.-Y. Pan, H. Yang, and C. Faloutsos. Mmss: Multi-modal story-oriented video summarization. In *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining*, 2004.

[19] Z. Pecenovic, H. Lausanne, P. F. Losanna, F. Dra, L. D. E. Lau, S. Anne, A. Dr, S. Ayer, and P. M. Vetterli. Image retrieval using latent semantic indexing, 1997.

[20] T.-T. Pham, N. E. Maillot, J.-H. Lim, and J.-P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *CIKM*, 2007.

[21] T.-T. Pham, N. E. Maillot, J.-H. Lim, and J.-P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *CIKM*, 2007.

[22] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *BMVC*, 2008.

[23] P.Quelhas, D.-P. Monay, J.-M. Odobez, and L. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005.

[24] E. H. R. Lienhart, S. Romberg. Multilayer plsa for multimodal image retrieval. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, 2009.

[25] M. S. Rainer Lienhart. Plsa on large scale image databases. In *ECCV '06: European Conference on Computer Vision*, 2006.

[26] Y. Rui, T. S. Huang, and S. fu Chang. Image retrieval: Past, present, and future. In *Journal of Visual Communication and Image Representation*, pages 1–23, 1997.

[27] B. Savas and L. Eldén. Handwritten digit classification using higher order singular value decomposition. *Pattern Recogn.*, 2007.

[28] Singh, P. M., Cunningham, and E. Curran. Active learning for multi-label image annotation. In *AICS*, 2008.

[29] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.

[30] 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.*, 2000.

[31] A. Smilde, R. Bro, and P. Geladi. Multi-way analysis: Applications in the chemical sciences. *Wiley, New York.*, 2004.

[32] C. G. Snoek. and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications.*, 2005.

[33] a. D. L. T.Landauer, P.Foltz. Introduction to latent semantic indexing. In *Discourse Processes*, 1998.

[34] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, 2002.

[35] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR*, 2008.

[36] X.-J. Wang, W.-Y. Ma, L. Zhang, and X. Li. Multi-graph enabled active learning for multimodal web image retrieval. In *MIR*, 2005.

[37] T. Yamaguchi and M. Maruyama. Feature extraction for document image segmentation by plsa model. In *DAS*, 2008.

[38] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieva. In *ICCV*, 2005.