A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieval

Ruofei Zhang[1], Zhongfei (Mark) Zhang[1], Mingjing Li[2], Wei-Ying Ma[2], Hong-Jiang Zhang[2]
[1]Dept. of Computer Science, SUNY at Binghamton, Binghamton, NY 13902, USA
[2]Microsoft Research Asia, Beijing 100080, China

Abstract

This paper addresses automatic image annotation problem and its application to multi-modal image retrieval. The contribution of our work is three-fold. (1) We propose a probabilistic semantic model in which the visual features and the textual words are connected via a hidden layer which constitutes the semantic concepts to be discovered to explicitly exploit the synergy among the modalities. (2) The association of visual features and textual words is determined in a Bayesian framework such that the confidence of the association can be provided. (3) Extensive evaluation on a large-scale, visually and semantically diverse image collection crawled from Web is reported to evaluate the prototype system based on the model. In the proposed probabilistic model, a hidden concept layer which connects the visual feature and the word layer is discovered by fitting a generative model to the training image and annotation words through an Expectation-Maximization (EM) based iterative learning procedure. The evaluation of the prototype system on 17,000 images and 7,736 automatically extracted annotation words from crawled Web pages for multi-modal image retrieval has indicated that the proposed semantic model and the developed Bayesian framework are superior to a state-of-the-art peer system in the literature.

1. Introduction

In traditional Content-Based Image Retrieval (CBIR), users have to provide examples of images that they are looking for. Similar images are found based on the match of image features. Even though there have been many studies on CBIR, empirical studies have shown that using image features solely to find similar images is usually insufficient due to the notorious *semantic gap* [19]. On the other hand, it is well-observed that often imagery does not exist in isolation; instead, typically there is rich collateral information co-existing with image data in many applications. Examples include the Web, many domain-archived image databases (in which there are annotations to images), and even consumer photo collections. In order to reduce the semantic gap, recently multi-modal approaches to image retrieval are proposed in the literature [22] to explicitly exploit the redundancy co-existing in the collateral information to the images. In addition to the improved retrieval accuracy, another benefit for the multi-modal approaches is the added querying modalities. Users can query an image database either by image, or by a collateral information modality (e.g., text), or by any combination.

In this paper, we propose a probabilistic semantic model and the corresponding learning procedure to address the problem of automatic image annotation and show its application to multi-modal image retrieval. Specifically, we use the proposed probabilistic semantic model to explicitly exploit the synergy between the different modalities of the imagery and the collateral information. In this work, we only focus on a specific collateral modality — text. The model may be generalized to incorporating other collateral modalities. Consequently, the synergy here is explicitly represented as a hidden layer between the image and the text modalities. This hidden layer constitutes the concepts to be discovered through a probabilistic framework such that the confidence of the association can be provided. An Expectation-Maximization (EM) based iterative learning procedure is developed to determine the conditional probabilities of the visual features and the words given a hidden concept class. Based on the discovered hidden concept layer and the corresponding conditional probabilities, the image-to-text and text-to-image retrieval are performed in a Bayesian framework.

In recent CBIR literature, Corel data have been extensively used to evaluate the retrieval performance [1, 8, 9, 15]. It has been argued [21] that the Corel data are relatively easy to annotate and retrieve due to its small number of concepts and small variations of visual contents. In addition, the relatively small number (1000 to 5000) of training images and test images typically used in the literature further makes the problem easier and the evaluation less convictive. In order to truly capture the difficulties in real scenarios such as Web image retrieval and to demonstrate the robustness and promise of the proposed model and framework in these challenging applications, we evaluate our prototype system on a collection of 17,000 images with the automatically extracted textual annotation from various crawled Web pages. To our knowledge, this scale of evaluation with this diversity has never been reported in the literature. We show that the proposed model and framework work well in this scale of noisy and diverse image data set and substantially

outperform the state-of-the-art peer system MBRM [9].

The rest of the paper is organized as follows: Section 2 discusses the related work on image annotation and multi-modal image retrieval. In Section 3 the proposed probabilistic semantic model and the EM based learning procedure are described. Section 4 presents the Bayesian framework developed to support the multi-modal image retrieval. The acquisition of the training and testing data collected from the Web, and the experiments to evaluate the proposed approach against a state-of-the-art peer system in several aspects are reported in Section 5. Finally the paper is concluded in Section 6.

2. Related Work

A number of approaches have been proposed in the literature on automatic image annotation [1, 8, 9, 15]. Different models and machine learning techniques are developed to learn the correlation between image features and textual words from the examples of annotated images and then apply the learned correlation to predict words for unseen images. The co-occurrence model [17] collects the co-occurrence counts between words and image features and uses them to predict annotated words for images. Barnard and Duygulu et al [1, 8] improved the co-occurrence model by utilizing machine translation models. The models are correspondence extensions to Hofmann's hierarchical clustering aspect model [12, 13, 11], which incorporate multi-modality information. The models consider image annotation as a process of translation from "visual language" to text and collect the co-occurrence information by the estimation of the translation probabilities. The correspondence between *blobs* and words are learned by using statistical translation models. As noted by the authors [1], the performance of the models is strongly affected by the quality of image segmentation. More sophisticated graphical models, such as Latent Dirichlet Allocator (LDA) [3] and correspondence LDA, have also been applied to the image annotation problem recently [2]. Another way to address automatic image annotation is to apply classification approaches. The classification approaches treat each annotated word (or each semantic category) as an independent class and create a different image classification model for every word (or category). One representative work of these approaches is automatic linguistic indexing of pictures (ALIPS) [15]. In ALIPS, the training image set is assumed well classified and each category is modeled by using 2D multi-resolution hidden Markov models. But it is notable that the assumption made in ALIPS that the annotation words are semantically exclusive is not valid in practice.

Recently, relevance language models [9] have been successfully applied to automatic image annotation. The essential idea is to first find annotated images that are similar to a test image and then use the words shared by the annotations of the similar images to annotate the test image. One model in this category is Multiple-Bernoulli Relevance Model (MBRM) [9], which is based on the Continuous-space Relevance Model (CRM) [14]. In MBRM, the word probabilities are estimated using a multiple Bernoulli model and the image block feature probabilities using a non-parametric kernel density estimate.

It has been noted that in many cases both images and word-based documents are interesting to users' querying needs, such as in the Web search environment. In these scenarios, multi-modal image retrieval, i.e., leveraging the collected textual information to improve image retrieval and to enhance users' querying modalities, is proven to be promising. Studies have been reported on this problem. Chang et al [5] applied Bayes point machine to associate words and images to support multi-modal image retrieval. In [23], latent semantic indexing is used together with both textual and visual features to extract the underlying semantic structure of Web documents. Improvement of the retrieval performance is reported attributed to the synergy of both modalities. Recently, approaches using multi-modal information for Web image retrieval are emerging. In [20], an iterative similarity propagation approach is proposed to explore the inter-relationships between Web images and their textual annotations for image retrieval. The mutual reinforcement of similarities between different modalities is exploited, which boosts the Web image retrieval performance.

3. Probabilistic Semantic Model

To achieve the automatic image annotation as well as multi-modal image retrieval, a probabilistic semantic model is proposed for the training image and the associated textual word annotation dataset. The probabilistic semantic model is developed by the EM technique to determine the hidden layer connecting image features and textual words, which constitutes the semantic concepts to be discovered to explicitly exploit the synergy between imagery and text.

First, a word about notation: $f_i, i \in [1, N]$ denotes the visual feature vector of images in the training database, where N is the size of the database. $w^j, j \in [1, M]$ denotes the distinct textual words in the training annotation word set, where M is the size of annotation vocabulary in the training database.

In the probabilistic model, we assume the visual features of images in the database, $f_i = [f_i^1, f_i^2, \ldots, f_i^L], i \in [1, N]$, are known i.i.d. samples from an unknown distribution. The dimension of the visual feature is L. We also assume that the specific visual feature annotation word pairs $(f_i, w^j), i \in [1, N], j \in [1, M]$ are known i.i.d. samples from an unknown distribution. Furthermore we assume that these samples are associated with an unobserved *semantic concept* variable $z \in Z = \{z_1, \ldots, z_K\}$. Each observation of one visual feature $f \in F = \{f_i, f_2, \ldots, f_N\}$ belongs to one or more concept classes z_k and each observation of one word $w \in V = \{w^1, w^2, \ldots, w^M\}$ in one image f_i belongs to one concept class. To simplify the model, we have two more assumptions. First, the observation pairs (f_i, w^j) are generated independently. Second, the pairs of random variables

(f_i, w^j) are conditionally independent given the respective hidden concept z_k,

$$P(f_i, w^j|z_k) = p_{\mathcal{F}}(f_i|z_k)P_{\mathcal{V}}(w^j|z_k) \quad (1)$$

The visual feature and word distribution is treated as a randomized data generation process described as follows: choose a concept with probability $P_{\mathcal{Z}}(z_k)$; select a visual feature $f_i \in F$ with probability $P_{\mathcal{F}}(f_i|z_k)$; and select a textual word $w^j \in V$ with probability $P_{\mathcal{V}}(w^j|z_k)$. As a result one obtains an observed pair (f_i, w^j), while the concept variable z_k is discarded.

Translating this process into a joint probability model results in the expression

$$\begin{aligned} P(f_i, w^j) &= P(w^j)P(f_i|w^j) \\ &= P(w^j)\sum_{k=1}^{K} P_{\mathcal{F}}(f_i|z_k)P(z_k|w^j) \end{aligned} \quad (2)$$

Inverting the conditional probability $P(z_k|w^j)$ in (2) with the application of the Bayes' rule results in

$$P(f_i, w^j) = \sum_{k=1}^{K} P_{\mathcal{Z}}(z_k)P_{\mathcal{F}}(f_i|z_k)P_{\mathcal{V}}(w^j|z_k) \quad (3)$$

The mixture of Gaussian [7] is assumed for the feature-concept conditional probability $P_{\mathcal{F}}(\bullet|Z)$. In other words, the visual features are generated from K Gaussian distributions, each one corresponding a z_k. For a specific semantic concept variable z_k, the conditional pdf of visual feature f_i is

$$p_{\mathcal{F}}(f_i|z_k) = \frac{1}{(2\pi)^{L/2}|\sum_k|^{1/2}} e^{-\frac{1}{2}(f_i-\mu_k)^T\sum_k^{-1}(f_i-\mu_k)} \quad (4)$$

where $\sum_k$ and μ_k are the covariance matrix and mean of visual features belonging to z_k, respectively. The word-concept conditional probabilities $P_{\mathcal{V}}(\bullet|Z)$, i.e., $P_{\mathcal{V}}(w^j|z_k)$ for $k \in [1, K]$, are estimated through fitting the probabilistic model to the training set.

Following the maximum likelihood principle, one determines $P_{\mathcal{F}}(f_i|z_k)$ by maximization of the log-likelihood function

$$\log \prod_{i=1}^{N} p_{\mathcal{F}}(f_i|Z)^{u_i} = \sum_{i=1}^{N} u_i \log(\sum_{k=1}^{K} P_{\mathcal{Z}}(z_k)p_{\mathcal{F}}(f_i|z_k)) \quad (5)$$

where u_i is the number of annotation words for image f_i. Similarly, $P_{\mathcal{Z}}(z_k)$ and $P_{\mathcal{V}}(w^j|z_k)$ can be determined by maximization of the log-likelihood function

$$\mathcal{L} = \log P(F, V) = \sum_{i=1}^{N}\sum_{j=1}^{M} n(w_i^j) \log P(f_i, w^j) \quad (6)$$

where $n(w_i^j)$ denotes the weight of annotation word w^j, i.e., occurrence frequency, for image f_i.

From (5), (6) and (2) we derive that the model is a statistical mixture model [16], which can be resolved by applying the EM technique [6]. Thus the probabilities can be iteratively determined by fitting the model to the training image database and the associated annotations.

Applying Bayes' rule to (3), we determine the posterior probability for z_k under f_i and (f_i, w^j):

$$p(z_k|f_i) = \frac{P_{\mathcal{Z}}(z_k)p_{\mathcal{F}}(f_i|z_k)}{\sum_{t=1}^{K} P_{\mathcal{Z}}(z_t)p_{\mathcal{F}}(f_i|z_t)} \quad (7)$$

$$P(z_k|f_i, w^j) = \frac{P_{\mathcal{Z}}(z_k)P_{\mathcal{Z}}(f_i|z_k)P_{\mathcal{V}}(w^j|z_k)}{\sum_{t=1}^{K} P_{\mathcal{Z}}(z_t)P_{\mathcal{F}}(f_i|z_t)P_{\mathcal{V}}(w^j|z_t)} \quad (8)$$

The expectation of the complete-data likelihood $\log P(F, V, Z)$ for the estimated $P(Z|F, V)$ derived from (8) is

$$\sum_{(i,j)=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M} n(w_i^j) \log [P_{\mathcal{Z}}(z_{i,j})p_{\mathcal{F}}(f_i|z_{i,j})P_{\mathcal{V}}(w^j|z_{i,j})]P(Z|F, V) \quad (9)$$

where $P(Z|F, V) = \prod_{s=1}^{N}\prod_{t=1}^{M} P(z_{s,t}|f_s, w^t)$. In (9) the notation $z_{i,j}$ is the concept variable that associates with the feature-word pair (f_i, w^j). In other words, (f_i, w^j) belongs to concept z_t where $t = (i, j)$.

Similarly, the expectation of the likelihood $\log P(F, Z)$ for the estimated $P(Z|F)$ derived from (7) is

$$\sum_{k=1}^{K}\sum_{i=1}^{N} \log(P_{\mathcal{Z}}(z_k)p_{\mathcal{F}}(f_i|z_k))p(z_k|f_i) \quad (10)$$

Maximizing (9) and (10) with Lagrange multipliers to $P_{\mathcal{Z}}(z_l)$, $p_{\mathcal{F}}(f_u|z_l)$, and $P_{\mathcal{V}}(w^v|z_l)$, respectively, under the following normalization constraints

$$\sum_{k=1}^{K} P_{\mathcal{Z}}(z_k) = 1, \sum_{k=1}^{K} P(z_k|f_i, w^j) = 1 \quad (11)$$

for any f_i, w^j and z_l, the parameters are determined as

$$\mu_k = \frac{\sum_{i=1}^{N} u_i f_i p(z_k|f_i)}{\sum_{s=1}^{N} u_s p(z_k|f_s)} \quad (12)$$

$$\sum_k = \frac{\sum_{i=1}^{N} u_i p(z_k|f_i)(f_i-\mu_k)(f_i-\mu_k)^T}{\sum_{s=1}^{N} u_s p(z_k|f_s)} \quad (13)$$

$$P_{\mathcal{Z}}(z_k) = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N} u(w_i^j)P(z_k|f_i, w^j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} n(w_i^j)} \quad (14)$$

$$P_{\mathcal{V}}(w^j|z_k) = \frac{\sum_{i=1}^{N} n(w_i^j)P(z_k|f_i, w^j)}{\sum_{u=1}^{M}\sum_{v=1}^{N} n(w_v^u)P(z_k|f_v, w^u)} \quad (15)$$

Alternating (7) and (8) with (12)–(15) defines a convergent procedure to a local maximum of the expectation in (9) and (10).

The number of concepts, K, is determined in advance for the EM model fitting based on the Minimum Description Length (MDL) principle [18] to maximize

$$\log(P(F, V)) - \frac{m_K}{2}\log(MN) \quad (16)$$

where the first term is expressed in (6) and m_K is the number of free parameters needed for a model with K mixture components. In our probabilistic model, we have

$$m_K = (K-1)+K(M-1)+K(N-1)+L^2 = K(M+N-1)+L^2-1$$

As a consequence of this principle, when models with different values of K fit the data equally well, the simpler model is selected.

4. Model based Image Annotation and Multi-modal Image Retrieval

After the EM-based iterative procedure converges, the model fitted to the training set is obtained. The image annotation and multi-modal image retrieval are conducted in a Bayeisan framework with the determined $P_{\mathcal{Z}}(z_k)$, $p_{\mathcal{F}}(f_i|z_k)$, and $P_{\mathcal{V}}(w^j|z_k)$.

Observing (1), the joint probability is

$$P(w^j, z_k, f_i) = P_{\mathcal{Z}}(Z_k)p_{\mathcal{F}}(f_i|z_k)P_{\mathcal{V}}(w^j|z_k) \quad (17)$$

Through applying Bayes law and the integration over $P_{\mathcal{Z}}(z_k)$, we obtain the following expression:

$$\begin{aligned} P(w^j|f_i) &= \int P_{\mathcal{V}}(w^j|z)p(z|f_i)dz \\ &= \int P_{\mathcal{V}}(w^j|z)\frac{p_{\mathcal{F}}(f_i|z)P(z)}{p(f_i)}dz \\ &= E_z\{\frac{P_{\mathcal{V}}(w^j|z)p_{\mathcal{F}}(f_i|z)}{p(f_i)}\} \end{aligned} \quad (18)$$

where

$$p(f_i) = \int p_{\mathcal{F}}(f_i|z)P_{\mathcal{Z}}(z)dz = E_z\{p_{\mathcal{F}}(f_i|z)\} \quad (19)$$

In above equations $E_z\{\bullet\}$ denotes the expectation over $P(z_k)$, the probability of semantic concept variables. (18) provides a principled way to determine the probability of word w^j for annotating image f_i. With the combination of (18) and (19), the automatic image annotation can be solved fully in the Bayesian framework.

In practice, we derive an approximation of the expectation in (18) by utilizing Monte Carlo sampling [10] technique. Applying Monte Carlo integration to (18) derives

$$\begin{aligned} P(w^j|f_i) &\approx \frac{\sum_{k=1}^{K} P_{\mathcal{V}}(w^j|z_k)p_{\mathcal{F}}(f_i|z_k)}{\sum_{h=1}^{K} p_{\mathcal{F}}(f_i|z_h)} \\ &= \sum_{k=1}^{K} P_{\mathcal{V}}(w^j|z_k)x_k \end{aligned} \quad (20)$$

where $x_k = \frac{p_{\mathcal{F}}(f_i|z_k)}{\sum_{h=1}^{K} p_{\mathcal{F}}(f_i|z_h)}$. The words with the top highest $P(w^j|f_i)$ are returned to annotate the image. Given this image annotation scheme, the image-to-text retrieval may be performed by retrieving documents for the returned words based on traditional text retrieval techniques.

Similar to the above derivation, the text-to-image retrieval is obtained by determining the conditional probability $P(f_i|w^j)$:

$$\begin{aligned} P(f_i|w^j) &= \int P_{\mathcal{F}}(f_i|z)P(z|w^j)dz \\ &= \int P_{\mathcal{V}}(w^j|z)\frac{p_{\mathcal{F}}(f_i|z)P(z)}{P(w^j)}dz \\ &= E_z\{\frac{P_{\mathcal{V}}(w^j|z)p_{\mathcal{F}}(f_i|z)}{P(w^j)}\} \end{aligned} \quad (21)$$

The expectation can be estimated as follows:

$$\begin{aligned} P(f_i|w^j) &\approx \frac{\sum_{k=1}^{K} P_{\mathcal{V}}(w^j|z_k)p_{\mathcal{F}}(f_i|z_k)}{\sum_{h=1}^{K} P_{\mathcal{V}}(w^j|z_h)} \\ &= \sum_{k=1}^{K} p_{\mathcal{F}}(f_i|z_k)y_k \end{aligned} \quad (22)$$

where $y_k = \frac{P_{\mathcal{V}}(w^j|z_k)}{\sum_h P_{\mathcal{V}}(w^j|z_h)}$. The images in the database with the top highest $P(f_i|w^j)$ are returned as the retrieval result for each query word.

5. Experiments

We have implemented the approach in a prototype system. The process of probabilistic model fitting to generate the image-concept-word model offline; the text querying and image querying based on the Bayesian framework are performed online. The system supports both image-to-text (i.e., image annotation) and text-to-image retrievals.

Two issues are taken consideration in the design of our experiments. First, the commonly used Corel database is relatively easy for image annotation and retrieval due to its limited semantics conveyed and small variations of visual contents. Second, the typical small scales of the datasets reported in the literature are far away from being realistic in all the real world applications. To address these issues, we decide not to use the Corel database in the evaluation of the prototype system; instead we evaluate the system on a collection of large-scale real world data automatically crawled from the Web. The images and the surrounding text describing the image contents in the Web pages are extracted from the blocks containing the images by using the VIPS algorithm [4]. The surrounding text is processed using the standard text processing techniques to obtain the annotation words. Apart from images and annotation words, the weight of each annotation word for images is computed by using a scheme incorporating TF, IDF, and the tag information in VIPS. The image-annotation word pairs are stemmed and manually cleaned in the training database for model fitting and testing. The data collection consists of 17,000 images and 7,736 stemmed annotation words. Among them, 12,000 images are used as the training set and the rest 5,000 images are used for the testing purpose. Compared with images in Corel, the images in this set are more diverse both on semantics and on visual appearance, which reflect the true nature of image search in many real applications.

The focus of this paper is not on image feature selection and our approach is independent of any visual features. For implementation simplicity and easy comparison purpose, similar features used in [9] are used in our prototype system. Specifically, a visual feature is a 36 dimensional vector, consisting of 24 color features (auto correlogram computed over 8 quantized colors and 3 Manhattan Distances) and 12 texture features (Gabor energy computed over 3 scales and 4 orientations).

To evaluate the effectiveness and the promise of the prototype system for multi-model image retrieval, the following performance measures are defined:

- Hit-Rate3 (HR3): the average rate of at least one word in the ground truth of a test image is returned in the top 3 returned words for the test set.
- Complete-Length (CL): the average minimum length of returned words which contains all the ground truth words for a

test image for the test set.

- Single-Word-Query-Precision (SWQP(n)): the average rate of relevant images (here 'relevant' means that the ground truth annotation of this image contains the query word) in the top *n* returned images for a single word query for the test set.

HR3 and CL measure the accuracy of image annotation (or the image-to-text retrieval); the higher the HR3, and/or the lower the CL, the better the annotation accuracy. SWQP(n) measures the precision of text-to-image retrieval; the higher the SWQP(n), the better the text-to-image retrieval precision.

Furthermore, we also measure the image annotation performance by using the annotation recall and precision defined in [9]. $recall = \frac{B}{C}$ and $precision = \frac{B}{A}$, where A is the number of images automatically annotated with a given word in the top 10 returned word list; B is the number of images correctly annotated with that word in the top-10-returned-word list; and C is the number of images having that word in ground truth annotation. An ideal image annotation system would have a high average annotation recall and annotation precision simultaneously.

Applying the method of estimating the number of hidden concepts to the training set, the number of the concepts is determined to be 262. Compared with the number of images in the training set, 12,000, and the number of stemmed and cleaned annotation words, 7,736, the number of semantic concept variables is far less. In terms of computational complexity, the model fitting is computation-intensive; it takes 45 hours to fit the model to the training set on a Pentium IV 2.3 GHZ computer with 1GB memory. Fortunately this process is performed offline and only once. For online image annotation and single-word image query, the response time is acceptable (less than 1 second).

To show the effectiveness and the promise of the probabilistic model in image annotation, we have compared the accuracy of our method with that of MBRM [9]. In MBRM, the word probabilities are estimated using a multiple Bernoulli model and no association layer between visual features and words is used. We compare our approach with MBRM because MBRM reflects the performance of the state-of-the-art automatic image annotation research. In addition, since the same image visual features are used in MBRM, a fair comparison of the performance is expected. Table 1 shows examples of the automatic annotation obtained by our prototype system and MBRM on the test image set. Here top 5 words (according to probability) are taken as automatic annotation of the image. It clearly indicates that our system performs noticeably better than MBRM.

The systematic evaluation results are shown for the test set in Table 2. Results are reported for all (7736) words in the database. Our approach clearly outperforms MBRM. As is shown, the average recall improves 48% and the average precision improves 69%. The multiple Bernoulli generation of words in MBRM is artificial and the association of the words and features is noisy. On the contrary, in our model no explicit word distribution is assumed and the synergy between the visual features and words exploited by the hidden concept variables reduces the noises substantially. We believe that these reasons account for the better performance of our approach.

Table 2: Performance comparison on the task of automatic image annotation on the test set.

Models	MBRM	Our Model
HR3	0.56	0.83
CL	1265	574
#words with $recall > 0$	3295	6078
Results on all 7736 words		
Average Per-word Recall	0.19	0.28
Average Per-word Precision	0.16	0.27

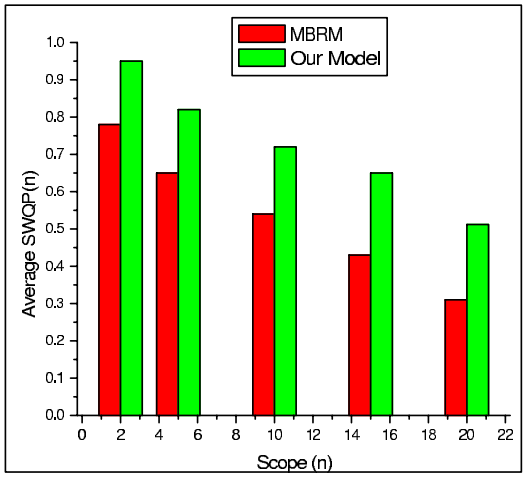


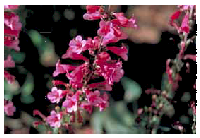
Figure 1: Average SWQP(n) comparisons between MBRM and our approach.

The single word text-to-image retrieval results on a set of 500 randomly selected query words are shown in Fig. 1. The average SWQP(2, 5, 10, 15, 20) of our system and those of MBRM are recorded. A returned image is considered as relevant to the single word query if this word is contained in the ground truth annotation of the image. It is shown that the performance of our probabilistic model has higher overall SWQP than that of MBRM. It is also noticeable that when the scope of the returned images increases the SWQP(n) in our system attenuates more gracefully than that in MBRM, which is another advantage of our model.

6. Conclusions

In this paper, we have developed a probabilistic semantic model for automatic image annotation and multi-modal image retrieval. Instead of assuming artificial distribution

Table 1: Examples of the automatic annotations produced by our prototype system and MBRM.

System					
MBRM	animal water wolf house tiger	male-face hair people bear sky	bird grass leopard sail cuckoo	flower red tree meadow outdoor	desert beach mummy building church
Our prototype system	wolf winter wild animal stone	male-face people hair man monologue	bird cuckoo yellow sand sky	flower red azalea leaf landscape	pyramid Egypt desert mummy beach

of annotation word and the unreliable association evidence used in many existing approaches, we assume a hidden concept layer as the connection between the visual features and the annotation words to explicitly exploit the synergy among the modalities. The hidden concept variables are discovered and the corresponding probabilities are determined by fitting the generative model to the training set. Based on the model obtained, the image-to-text and text-to-image retrieval are conducted in a Bayesian framework. The proposed model is promising for image annotation and multi-model image retrieval, which are demonstrated by the evaluation of the prototype system on 17,000 images and the automatically extracted annotation words from crawled Web pages. In comparison with a state-of-the-art image annotation system, MBRM, higher reliability and the superior effectiveness of our model and retrieval framework are reported.

7. Acknowledgement

This work is supported in part by Microsoft Research Asia through a graduate student internship and a faculty visiting researchership. Zhen Guo of SUNY Binghamton helped in part in the evaluations.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] D. Blei and M. Jordan. Modeling annotated data. In *the 26th International Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[3] D. Blei, A. Ng, and M. Jordan. Dirichlet allocation models. In *The International Conference on Neural Information Processing Systems*, 2001.

[4] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: a vision-based page segmentation algorithm. Microsoft Technical Report (MSR-TR-2003-79), 2003.

[5] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(1), January 2003.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society*, 39(1):1–38, 1977.

[7] W. R. Dillon and M. Goldstein. *Multivariate Analysis, Mehtods and Applications*. John Wiley and Sons, New York, 1984.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *The 7th European Conference on Computer Vision*, volume IV, pages 97–112, Copenhagon, Denmark, 2002.

[9] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *The International Conference on Computer Vision and Pattern Recognition*, Washington, DC, June, 2004.

[10] G. Fishman. *Monte Carlo Concepts, Algorithms and Applications*. Springer Verlag, 1996.

[11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[12] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. *AI Memo*, 1625, 1998.

[13] T. Hofmann, J. Puzicha, and M. I. Jordan. Unsupervised learning from dyadic data. In *The International Conference on Neural Information Processing Systems*, 1996.

[14] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *the International Conference on Neural Information Processing Systems (NIPS'03)*, 2003.

[15] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(9), September 2003.

[16] G. Mclachlan and K. E. Basford. *Mixture Models*. Marcel Dekker, Inc., Basel, NY, 1988.

[17] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on difiding and vector quantizing images with words. In *the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[18] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[19] 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. on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[20] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *the 12th annual ACM international conference on Multimedia*, pages 944–951, New York City, NY, 2004.

[21] T. Westerveld and A. P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *the SIGIR Multimedia Information Retrieval Workshop 2003*, August 2003.

[22] Z. M. Zhang, R. Zhang, and J. Ohya. Exploiting the cognitive synergy between different media modalities in multimodal information retrieval. In *the IEEE International Conference on Multimedia and Expo (ICME'04)*, Taipei, Taiwan, July 2004.

[23] R. Zhao and W. I. Grosky. Narrowing the semantic gap — improved text-based web document retrieval using visual features. *IEEE Trans. on Multimedia*, 4(2), 2002.