Semantic Image Retrieval Using Correspondence Topic Model with Background Distribution

Nguyen Anh Tu
Department of Computer Engineering
Kyung Hee University, Korea
Email: tunguyen@oslab.khu.ac.kr

Jinsung Cho
Department of Computer Engineering
Kyung Hee University, Korea
Email: chojs@khu.ac.kr

Young-Koo Lee
Department of Computer Engineering
Kyung Hee University, Korea
Email: yklee@khu.ac.kr

***Abstract*—Social image search becomes an active research field in recent years due to the rapid development in big data processing technologies. In the retrieval systems, text description/tags play a key role to bridge the semantic gap between low-level features and higher-level concepts, and so guarantee the reliable search. However, in practice manual tags are usually noisy and incomplete, resulting in a limited performance of image retrieval. To tackle this problem, we propose a probabilistic topic model to formalize the correlation of tags with visual features via the latent semantic topics. Our proposed approach allows us to effectively annotate and refine tags based on a Monte Carlo Markov Chain algorithm for approximate inference. Moreover, we present a measuring scheme using the refined tags and extracted topics for ranking the images. The experimental results from two large benchmark datasets show that our approach provides promising accuracy.**

***Index Terms*—Topic modeling, probabilistic graphical model, tag refinement, image annotation, image retrieval**

I. Introduction

Nowadays emerging social websites (e.g. Facebook, Flickr, Youtube, and Picasa) have quickly become a powerful part of Internet with over billions images uploaded by users. Particularly, the amount of stored multimedia data is growing rapidly due to the prevalence of digital cameras. Retrieving information, especially digital images, in such huge data poses great challenges, because there exists various types of information such as text, image feature, and user. In content-based image retrieval [1], [2], most methods highly depend on visual features like color or textures to derive a representation of image contents for estimating visual similarity between query and database images. These methods are very efficient for indexing with large-scale database, but their performance usually limited due to the semantic gap between low-level features and higher-level concept of each image. Moreover, semantic search with query provided in natural language now attracts a lot of attention from information retrieval and data mining community. In this approach, the semantic content of image is described by manual tags. However, it is extremely time-consuming to annotate in a huge database. Otherwise, the reliability of manual tagging is not guaranteed, because tags based on user's objectivity can be noisy, incomplete, and irrelevant.

To address the issue of manual tagging, there are significant efforts to design automatic annotation systems for inferring and refining the tags. Most of conventional works [3], [4], [5] treat image annotation as a prediction or classification task, where the classifiers are learned from the training data to map low-level features into relevant tags. In [4], the authors apply a multi-label classification to explore the relation between multiple labels. The major disadvantage of classification-based approaches is that its performance is limited to small vocabulary of tags with well-labeled training data. This is very difficult to apply in practice, because most images from social websites contain noisy tags with large vocabulary. Otherwise, machine learning techniques can also be adopted to automatic annotation like in [6], [7], [8]. For example, tag propagation in [6] is performed by using a weighted combination of keyword presence and absence among neighbors. Despite obtaining good performance, these approaches also require a large number of well-annotated training images like classification based approaches. On the other hand, statistical generative models [9], [10], [11], [12] have been employed to study the correlation between image and text description by using common latent variables. In general, these approaches scale well to database size and the number of tags. Particularly, the Correspondence LDA (CorrLDA) model in [9], [12] provides a natural way to learn latent topics from text word and image features. This allows us to encode human knowledge as well as deal with synonyms and homonyms in annotation. However, most of methods focus on employing CorrLDA for image annotation task rather than retrieval task as our interest here.

In this paper, we propose a retrieval framework that enables not only the effective image annotation and tag refinement but also an efficient similarity measurement for ranking retrieved images. Each image is modeled as a set of local features (e.g. SIFTs [13]), which typically achieves invariance of orientation and scale in modern visual recognition. These features are quantized to form a vocabulary of visual words. We then propose a probabilistic topic model to extract the semantic topics from the co-occurrence of tags and image content in term of "bag-of-visual words". Our proposed topic model called Correspondence Topic Model with Background Distribution (CTMB) is an extension of CorrLDA with consideration of background words. More specifically, visual words often appearing in the same images intend to have the same topic or belong to a specific object, which can be used to generate the corresponding tags. Hence, we can effectively complete the

Fig. 1. Example of a resulting image after applying to CTMB model. Image contains visual words which are labeled by different colors corresponding to different topics. Each topic can be referred as an object. The italic word "highway" indicates a noisy tag.

missing tags through extracted topics. Furthermore, our model also discovers background words which frequently appear in database and have a negative effect to topic extraction. These words should be removed during retrieval process to improve the accuracy. Fig. 1 shows the example of a resulting image after applying to our CTMB model. Benefiting from the extracted topics and refined (or complete) tags, we present an efficient scheme to estimate the similarity between query and database images. While previous works [9], [6], [7] only use predicted tags to compute similarity and may lead to unsatisfying results, our approach combine both refined tags and extracted topic to achieve more robust performance. Another advantage of using CTMB in our retrieval framework is that we are able to perform many types of queries for searching such as image, keywords, and combination of both.

Our major contributions are two-fold: (1) We propose a generative topic model built on human knowledge to formulate the correlation of visual words, text words, and background. This allows us to predict missing tags for refining task. (2) We present a scoring scheme to efficiently compute the similarity between two images. Using this scheme, proposed approach shows very promising results on public datasets.

The remainder of this paper is organized as follows. Section 2 introduces an overview of proposed framework and preprocessing step of image data. Section 3 describes our proposed topic model with parameter estimation and inference procedure, and then presents the approach of similarity measurement. Experimental results on standard datasets are conducted and discussed in Section 4. Conclusions are presented in Section 5.

II. Overview of Retrieval Framework

In this section, we present an overview of our retrieval framework as shown in Fig. 2. It also describes how the proposed approaches come together during the retrieval process. First, we preprocess each image, which can be the query or social image with manual tags in the database, to construct the preliminary representation. We then extract the regions of interest from the image via an affine invariant detector. The detected regions (or image patches) are described using 128-D SIFT descriptor. Each descriptor of the image is quantized [1] to a visual word via visual vocabulary or codebook learned by k-means clustering. Thus, each image can be represented by two types of entity: bag-of-visual words (BoV) and text words (tags). As illustrated in Fig. 2, our framework composed of an offline process and online process. The offline process extracts topics from BoV of each database image to generate corresponding tags. During this stage, the model parameters of CTMB model are learned from visual words and manually annotated tags of social images. The online process aims to retrieve the ranked list of relevant images based on the extracted topics and refined tags of query and database images. In this stage, CTMB use parameters estimated from learning process to speedup the querying time and so improve the scalability of our method. It should be note that the query can be image, keywords, or combination of both. If the query is only keywords, it can be directly measured with the refined tags of database images without applying to CTMB model.

III. Probabilistic Topic Model for Image Retrieval

In this section, we introduce the proposed generative latent topic model including parameter estimation and inference procedure.By taking advantage of topic modeling, we further propose a method for similarity measurement by fully considering the information associated with semantic content.

A. Correspondence Topic Model with Background Distribution (CTMB)

In topic modeling, we treat an image as the combination of visual words and text words, and can directly apply to topic models dealing with multiple type entity data like CorrLDA model to learn or infer common topic. However, CorrLDA model (as shown in Fig. 3(a)) has not been fully exploited for retrieval task in previous works [9], [12], where they only focus on annotation and classification task. Unlike conventional topic models, we incorporate explicit notions strongly related to image retrieval into our model. Particularly, inspired by [14], our topic model considers the background distribution of visual words because we need to remove visual words having low semantic meaning. In a retrieval task, this type of visual word, which appear in almost all images, is meaningless for similarity measurement.

Our proposed topic model is illustrated by the graphical model shown in Fig. 3(b). We can see that our model is an extension of CorrLDA by incorporating background information. The CTMB model is learned in an unsupervised manner. We summarize a formal description of the CTMB in Table I.

The CTMB model represents a collection of D images, and each image I_d consists of N_v visual words and N_w text words. We use latent variables (i.e., z_{di}) to characterize the topics, where topic z for each visual feature indicates which

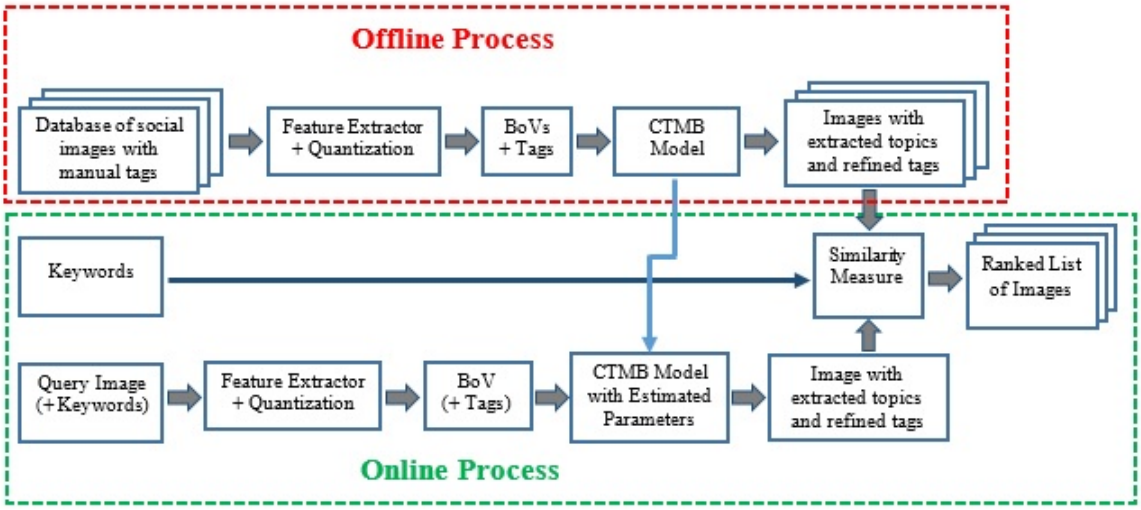


Fig. 2. Overview of proposed framework

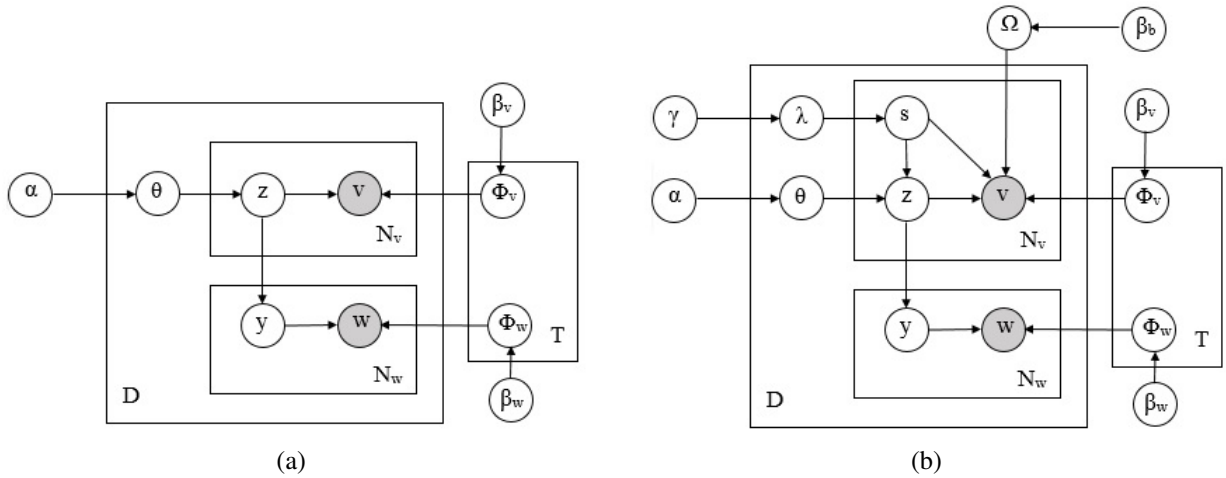


Fig. 3. Probabilistic topic model: (a) Correspondence LDA (CorrLDA), (b) Correspondence Topic Model with Background Distribution (CTMB).

TABLE I
NOTATION OF PARAMETERS IN CTMB

Parameters	Description
$\alpha, \beta_v, \beta_w, \beta_b, \gamma$	Dirichlet hyper parameters
$\lambda_d \in \mathbb{R}^2$	Bernoulli distribution of word type (i.e., visual topic and background) in image d
$\Omega \in \mathbb{R}^V$	Mult distribution of visual words in background
$\theta_d \in \mathbb{R}^T$	Mult distribution of topics in image d
$\phi_{v,t} \in \mathbb{R}^V$	Mult distribution of visual words in topic t
$\phi_{w,t} \in \mathbb{R}^W$	Mult distribution of text words in topic t

objects or parts of object it comes from. More specifically, we factor the image into a combination of T topics. Each topic is modeled as two distributions over visual vocabulary of size V and over textual vocabulary of size W. The textual topic (denoted by y variable) is a counterpart of topic z. Thus, CTMB model directly use the latent topic of visual words for generating the text words. According to the graphical model, v_{di} and w_{di} are the only observable variables, and the others are latent variables. Formally, the generative process of our CTMB model for image corpus is as follows:

1) For each topic t:
   a) Draw an appearance distribution $\phi_{v,t} \sim Dir(\beta_v)$
   b) Draw an textual distribution $\phi_{w,t} \sim Dir(\beta_w)$
2) Draw background distribution $\Omega \sim \text{Dir}(\beta_b)$
3) For each image $I_d(d = 1, ..., D)$:
   a) Draw word type distribution $\lambda_d \sim \text{Beta}(\gamma)$
   b) Draw topic proportion $\theta_d \sim \text{Dir}(\alpha)$
4) For each visual word v_{di} where $i \in 1, 2, ..., N_v$:
   a) Draw switch sample $s_{di} \sim \text{Bernoulli}(\lambda_d)$
   b) if $s_{di} = 1$
      i) Draw topic $z_{di} \sim \text{Multi}(\theta_d)$
      ii) Draw visual word $v_{di} \sim \text{Multi}(\phi_{v,z_{di}})$

c) if $s_{di} = 2$

i) Draw visual word $v_{di} \sim \text{Multi}(\Omega)$

5) For each text word w_{dj} where $j \in 1, 2, ..., N_w$:

a) Draw topic $y_{dj} \sim \text{Unif}(z_1, ..., z_{Nv})$

b) Draw text word $w_{dj} \sim \text{Multi}(\phi_{w,y_{dj}})$

Here, Dir and Multi denote Dirichlet and Multinomial distributions, respectively. The prior Multi is chosen to conjugate to Dir for the word distributions, and hence, they simplify computation and guarantee efficient inference. We can see that switch variable s is used to control the generation of the visual word. An image contains two types of visual words, where one is generated from topic distribution Multi(Φ_v) and the other is generated from background distribution Multi(Ω), leading to direct correlation between the visual words and background. Otherwise, the topic y corresponds to one of the visual topic z's through uniform distribution, and text word is then generated from topic distribution Multi(Φ_w). Therefore, the correlation of visual and text words is highly enforced using this model.

B. Parameter estimation in CTMB

In this subsection, we describe a method for parameter estimation in the CTMB model where we will use an training set of D images. Let $\Pi = \{\alpha, \beta_v, \beta_w, \beta_b, \gamma, \}$ be the set of hyper parameters. Given a corpus of image data $\{\mathbf{v_d}, \mathbf{w_d}\}_{d=1}^{D}$, the model parameters $\Phi_v, \Omega \in \mathbb{R}^{V \times T}$, and $\Phi_w \in \mathbb{R}^{W \times T}$ of visual topic, background and textual topic distributions respectively, can be found by maximization of the following log likelihood function.

$$L(\Phi_v, \Phi_w, \Omega) = \sum_{d=1}^{D} \log\left(p(\mathbf{v_d}, \mathbf{w_d}, \mathbf{z_d}, \mathbf{y_d}, \mathbf{s_d} | \Phi_v, \Phi_w, \Omega, \Pi)\right) \tag{1}$$

The distribution in Eq. 1 is intractable to direct estimation, so one effective approach is to estimate using Monte Carlo EM algorithms [15], as summarized in **Algorithm 1**.

Algorithm 1 Parameter estimation of CTMB

Input: Corpus of image data formed as a bag of visual words and text words $\{\mathbf{v_d}, \mathbf{w_d}\}_{d=1}^{D}$

Output: The estimated parameters Φ_v, Φ_w and Ω

1) **Initialization**. Initialize set of parameters $\left\{\Phi_v^{(0)}, \Phi_w^{(0)}, \Omega^{(0)}\right\}$

2) **For each** $k = 1, ..., K$ **do**:

a) Given $\left\{\Phi_v^{(k-1)}, \Phi_w^{(k-1)}, \Omega^{(k-1)}\right\}$, sample latent variables with N Gibbs steps for each image I_d from the posterior distribution using Eqs. 2,3,4.

b) Compute $\left\{\Phi_v^{(k)}, \Phi_w^{(k)}, \Omega^{(k)}\right\}$ using as Eqs. 6,7,8.

3) **End**

As shown in Algorithm 1, rather than directly computing the posterior of latent variables, which is intractable, we draw samples from it. Then, the parameters are estimated by examining this posterior distribution. Here, we use the collapse Gibbs sampling algorithm [16] for joint sampling of latent variables z, s relevant to visual word v, and sampling of latent variable y relevant to text word w, as in the following equations:

$$p(z_{di} = t, s_{di} = 1 | \mathbf{v_d}, \mathbf{z_{-di}}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{v1,-di} + \gamma}{N_{v,-di} + 2\gamma} \times \frac{n_{vt,-di}^{VT} + \beta_v}{\sum_{v'} n_{v't,-di}^{VT} + V\beta_v} \times \frac{n_{td,-di}^{TD} + \alpha}{\sum_{t'} n_{t'd,-di}^{TD} + T\alpha} \tag{2}$$

$$p(s_{di} = 2 | \mathbf{v_d}, \mathbf{s_{-di}}, \Pi) \propto \frac{N_{v2,-di} + \gamma}{N_{v,-di} + 2\gamma} \times \frac{n_{v,-di}^{V} + \beta_b}{\sum_{v'} n_{v',-di}^{V} + V\beta_b} \tag{3}$$

$$p(y_{dj} = t | \mathbf{w_d}, \mathbf{z}, \mathbf{y_{-di}}, \Pi) \propto \frac{n_{wt,-dj}^{WT} + \beta_w}{\sum_{w'} n_{w't,-di}^{WT} + W\beta_w} \times \frac{n_{td}^{TD}}{N_v} \tag{4}$$

where the subscript $-di$ indicates whole variables excluding the i^{th} variable in image d. N_{v1} and N_{v2} are the numbers of visual words in image d assigned to the related topic and background words, respectively; n_{td}^{TD} is the number of visual words assigned to topic t in image d; n_{vt}^{VT} is the number of times word v is assigned to topic t; n_v^V is the number of times word v is assigned to the background words distribution in the image corpus; n_{wt}^{WT} is the number of times text word w is assigned to topic t.

The above equations are obtained by marginalizing over parameters Φ_v, Φ_w, Ω, and θ separately. One can observe that the first terms of Eqs. 2 and 3 indicate the ratio of visual words assigned to the topic distribution ($s_{di} = 1$) and background distribution ($s_{di} = 2$), respectively. The second term of Eq. 2 represents the probability of visual word v_{di} under topic t, whereas the second term of Eq. 3 represents the probability of background words. The last term of Eq.2 shows the probability of topic t in image d. Moreover, Eq. 4 measures the probability that the observed text word w_{di} is assigned to topic t, where its last term indicates the correspondence with visual content via proportion of visual words assigned to topic t in one image.

Since all latent variables are computed from sampling equations, parameters Φ_v, Φ_w, Ω are then estimated by examining posterior distributions. Following some iterative steps, the parameters will converge to $\Phi_v^*, \Phi_w^*, \Omega^*$. The posterior of the topic-visual word multinomial is computed as belows:

$$p(\Phi_{v,t} | \mathbf{v}, \mathbf{z}, \mathbf{s}) = \text{Dir}\left\{\beta_v + n_{vt}^{VT}\right\} \tag{5}$$

where $\mathbf{v} = \{\mathbf{v_d}\}_{d=1}^{D}$, $\mathbf{z} = \{\mathbf{z_d}\}_{d=1}^{D}$, $\mathbf{s} = \{\mathbf{s_d}\}_{d=1}^{D}$. Thus, Φ_v can be estimated as the posterior mean of $p(\Phi_{v,t} | \mathbf{v}, \mathbf{z}, \mathbf{s})$, which is simply the normalized Dirichlet parameters, as follows:

$$\Phi_{v,t} = \frac{n_{vt}^{VT} + \beta_v}{\sum_{v'} n_{v't}^{VT} + V\beta_v} \tag{6}$$

Similarly, we can estimate Ω and Φ_w of the background and textual topic distributions as follows:

$$\Omega = \frac{n_v^V + \beta_b}{\sum_{v'} n_{v'}^V + V\beta_b} \quad (7)$$

$$\Phi_{w,t} = \frac{n_{wt}^{WT} + \beta_w}{\sum_{w'} n_{w't}^{WT} + W\beta_w} \quad (8)$$

C. Inference of unseen image and tag correspondence

An unseen image (e.g. a query, database image) can be applied to our model. With known parameters (i.e. Φ_v, Φ_w, Ω) obtained from the training process, we infer the latent variables of the unseen image, such as the topic labels z_{di} and y_{dj}. By using CTMB, the inference algorithm is similar to the estimation. However, the second terms in Eqs. 2 and 3 will be fixed and replaced by $\Phi_{v,t}$ and Ω, respectively, while the first term in Eq. 4 will be replaced by $\Phi_{w,t}$. This reflects the fact that all the learned object and background could be present without any prior knowledge. Note that the inference of different images are independent of each other, and the update equation of the Gibbs sampling for inference can be factored into the terms that only depend on variables related to a single image. Therefore, we can distribute images to multiple machines and process them in parallel. Hence, there is no issue with the scalability of our approach.

In CTMB model, the correspondence of a tag w for each image I_d is formulated as the probability conditioned on the set of image features (or visual words) $\mathbf{v_d}$. It can be computed as follows:

$$P(w|\mathbf{v_d}) = \sum_t P(w|t)P(t|\mathbf{v_d}) = \sum_t \Phi_{w,wt}\theta_{td} \quad (9)$$

Hence, tag prediction for image I_d can be performed by a dot product between the w^{th} row of matrix of tag-topic distribution Φ_w and d^{th} column of matrix of topic-document proportion θ. The most relevant tags are ranked based on the computed probability. Eq. 9 also shows how our approach deal with the tag refinement. More specifically, the probability of irrelevant tags should be small, whereas the probability of missing tags are increased via the extracted topics.

D. Semantic image retrieval

Given the predicted or refined tags, previous text-based approaches only use tag information for the retrieval task. This may lead to unsatisfying results if the annotation is incorrect or too general. For example, image annotated by the tag "animal" can be retrieved by different images of "dog", "cat", or "tiger" with very low visual similarity. This limitation motivates us to combine the visual and textual information to achieve more reliable search. In this work, the visual content information is represented by topic-document proportion θ which is strongly related to the generation of visual words. Given a query image I_q and database image I_d, their visual representations are θ_q and θ_d. Let r_q and r_d be two W-dimensional vectors representing for textual information or semantic content of images I_q and I_d, respectively, where w^{th} element of these vectors correspond to the probability of refined tag w computed by Eq. 9. Then, the similarity score between two images is defined as follows:

$$S(q,d) = (1-\mu)\theta_q\theta_d + \mu r_q r_d \quad (10)$$

where the parameter $\mu \in [0,1]$ controls the weight of textual similarity in the above score. We could set this value based on user preference or the type of query. Subsequently, we rank all database images according to the final scores and return the most relevant images to the user. Particularly, if the query is keywords, then we can simply apply it to scoring stage by setting $\mu = 1$. In this case, $r_q \in \{0,1\}^W$ where its element i^{th} is set to 1 if the i^{th} tag appear in the query. When the query is an image or image with keywords, the visual representation θ and textual representation of refined tags r are estimated after applying to CTMB model. In this paper, we set $\mu = 0.5$ for this case, because we consider the equal importance of two kind of information. It should be note that if the query is an image without keywords, the text part of CTMB model is excluded and the topics are purely extracted from the visual features.

TABLE II
STATISTICS OF BENCHMARK DATASETS

	Labelme	IAPR TC12
Number of images	2920	19805
Vocabulary size of tags (W)	490	291
Tags per image (mean/max)	11/48	5.7/23

IV. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed approach using CTMB model by comparing with CorrLDA model and other state-of-the-art methods on public benchmark datasets for image retrieval. We also studied the topic learning by examining the average log-likelihood of proposed model. Our results are reported for evaluating the both qualitative and quantitative performance.

Datasets. Two benchmark datasets, as summarized in Table II, are used in this paper:

- Labelme [17]: It contains 2920 online photos, manually annotated by 490 noun tags corresponding to the objects and object classes. The maximum number of annotated tags per images is 48.
- IAPR TC12 [6]: it consists of 19805 still natural images taken from locations around the world. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes. The vocabulary of popular tags contains 291 words. The maximum number of annotated tags per images is 23.

Experimental setup. For all datasets, the vocabulary size V of visual words was set to 2000. The hyper parameters of the CTMB model are set as following: $\alpha = 0.2, \beta_b = \beta_v = 0.01, \beta_w = 0.1$, and $\gamma = 0.5$. The values of these parameters are empirically selected to obtain the best results. To quantitatively evaluate the retrieval performance, we used

mean average precision (MAP) for comparison with the competing methods. The retrieval performance of a single query was measured by the average precision (AP), which is the area under the precision recall curve. Subsequently, the mean value over multiple queries was the final measurement of the retrieval performance. For the keyword queries used in our experiments, we collect them from the textual vocabularies of each dataset. Each query may contain a single tag or a combination of tags.

A. Topic learning

We can evaluate the performance of topic model when it was estimated during offline process from the training data. The estimation of the topic model is obtained by running Gibbs Sampling until convergence. As shown in Fig. 4(a), CTMB takes about 50 iterations to converge for model estimation. Here, log-likelihood is used for evaluation, which reflects the fitting of topic model with the training data. The higher score of log-likelihood is better. The marginal likelihood $P(\mathbf{v}|\mathbf{z})$ can be computed by integrating out latent variables as follows:

$$P(\mathbf{v}|\mathbf{z}) = \left(\frac{\Gamma(V\beta_v)}{\Gamma(\beta_v)^V}\right)^T \times \prod_{t=1}^{T} \frac{\prod_v (n_{vt}^{VT} + \beta_v)}{\Gamma\left(\sum_{v'} n_{v't}^{VT} + V\beta_v\right)} \times \frac{\Gamma(V\beta_b)}{\Gamma(\beta_b)^V} \times \frac{\prod_v (n_v^V + \beta_b)}{\Gamma\left(\sum_{v'} n_{v'}^V + V\beta_b\right)} \quad (11)$$

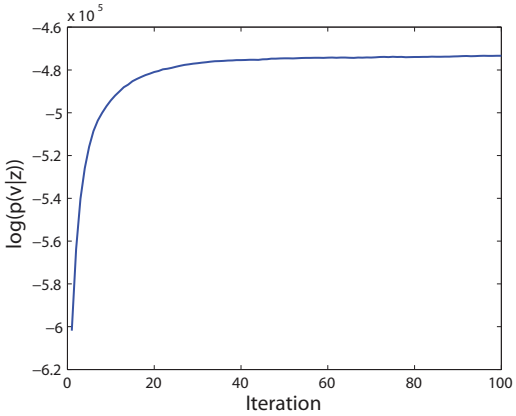
In CTMB model, the number of topics T is the free parameter. To select this number empirically, we examined the effect of this parameter on the log-likelihood of the CTMB model, as shown in Fig. 4(b). We achieve the maximum likelihood with about 130 topics, and after that, the likelihood slightly decreases with increasing T. This is because the variety of content in the images in the dataset makes the performance better with higher rather than lower topics. But if T is too high, it will degrade the content of image data. In this study, we chose T = 130 as the optimal number of topics for our remaining experiments.

B. Performance comparison and evaluation

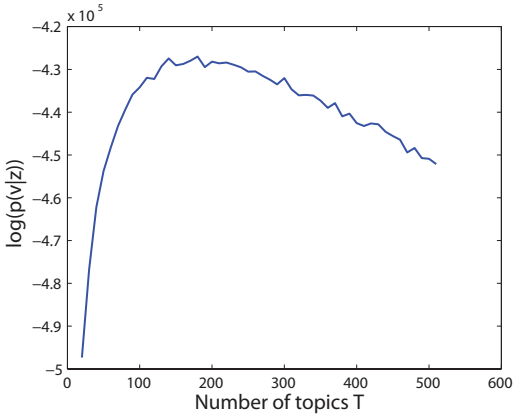
The improvement of performance is investigated using our proposed CTMB model. We compared our methods with CorrLDA and a number of well-known methods based on automatic annotation including Tag Relevance by Neighborhood Voting (TagNV) [7] and Tag propagation (TagProp) [6].

In the first experiment, we compare our performance with other methods by using the keyword queries, which can be single tag or multiple tags. Table III shows the MAP results for different datasets. We can observe that CTMB outperforms significantly TagNV, TagProp, and CorrLDA in all cases. Otherwise, the MAP of multiple tags is consistently lower than MAP of single tag due to its higher complexity.

Furthermore, we conducted other experiment using image query. It is closely related to the query of multiple keywords, because we will perform tag prediction on query image before measuring its similarity with database images. In this experiment, our method uses two values of $\mu = 1$ and $\mu = 0.5$ as



(a)



(b)

Fig. 4. Evaluation of topic learning process : (a) Log-likelihood over the iterations, (b) Impact of the topic number.

TABLE III
MAP FOR IMAGE RETRIEVAL USING KEYWORD QUERY

Methods	TagNV	TagProp	CorrLDA	CTMB
Labelme (Single)	0.742	0.755	0.728	0.771
IAPR TC12 (Single)	0.622	0.614	0.593	0.655
Labelme (Multiple)	0.676	0.654	0.684	0.713
IAPR TC12 (Multiple)	0.568	0.562	0.547	0.593

presented in Eq. 10, where $\mu = 1$ corresponds to purely textual representation and $\mu = 0.5$ corresponds to combination of textual and visual representation. We can observe from Table **??** that CTMB also outperforms other methods significantly. In addition, the performance of combined representation ($\mu = 0.5$) is better than textual representation ($\mu = 1$). This shows the improvement of our scoring scheme compared to traditional similarity measurement.

To evaluate the computational efficiency, we performed the third experiment based on the running time of each method. Here, the experiments were run with Matlab on an Intel Core i5 @2.5GHz and 8GB RAM. Table V summarizes the running time of all methods in comparison, where it shows the average time cost for a test image. We observe that although

TABLE IV
MAP FOR IMAGE RETRIEVAL USING IMAGE QUERY

Methods	TagNV	TagProp	CorrLDA	CTMB ($\mu = 1$)	CTMB ($\mu = 0.5$)
Labelme	0.691	0.703	0.695	0.707	0.724
IAPR TC12	0.564	0.551	0.555	0.584	0.617

TABLE V
COMPARISON OF AVERAGE TIME COST (IN SECONDS) FOR DIFFERENT METHODS

Methods	TagNV	TagProp	CorrLDA	CTMB
Labelme	0.044	0.021	0.051	0.047
IAPR TC12	0.028	0.016	0.035	0.032

CTMB is slower than TagProp, it is comparable to TagNV and slightly faster than CorrLDA. Otherwise, the training time for parameter estimation during offline process is 250 seconds on Labelme dataset, and 2200 seconds on IAPR TC12 dataset.

The qualitative results of the proposed approach is shown in Fig. 5, where we use two types of query consisting of image and image combined keywords. We can see that the retrieval results for queries of image combined keywords (upper row) are more relevant than the image queries (lower row). This is because the retrieved results of query with only image intend to reflect the extracted topics from the query. For example, in the second query, topics related to "tree" and "building" dominate other topics (or objects). It leads to that images containing these dominated objects have higher probability to appear in ranked list. For query of image combined keywords, most results are highly relevant to the queries in both visual and semantic content. Thus, in reality, when applying our proposed method, we can use a query image with additional keywords for refining search results.

V. CONCLUSION

In this paper, we have presented a framework for semantic image retrieval in large-scale database. We first introduced a probabilistic topic model built in the correlation between tags and visual features to extract the latent semantic topics by performing Gibbs sampling algorithm for approximate inference. We also integrated background information into topic model to improve the accuracy of tag prediction, tag refinement, and image retrieval task . Finally, we proposed a scoring scheme which is the combination of semantic and visual representation via refined tags and extracted topic to overcome the limitation of traditional similarity measure. The experimental results verified that our CTMB model behaved very well on social image data and achieved better performance than the well-known methods for refining tags and image retrieval. By taking the advantage of topic model, our method can perform multiple types of query and obtain the promising search results.

ACKNOWLEDGMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-(H8501-15-1015) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[3] W. Jiang, S.-F. Chang, and A. C. Loui, "Active context-based concept fusionwith partial user labels," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2917–2920.

[4] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 17–26.

[5] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 7, pp. 1281–1294, 2011.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 309–316.

[7] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *Multimedia, IEEE Transactions on*, vol. 11, no. 7, pp. 1310–1322, 2009.

[8] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He, "Mining social images with distance metric learning for automated image tagging," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 197–206.

[9] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.

[10] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: constraining the latent space," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 348–351.

[11] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 394–410, 2007.

[12] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1903–1910.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] C. Chemudugunta and P. S. M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, vol. 19. MIT Press, 2007, p. 241.

[15] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[16] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

Fig. 5. Qualitative results by using the queries of image and image combined single or multiple keywords. The queries are outlined by blue color, while the irrelevant images are outlined by red color. Each query corresponds to two rows where the upper row is for image query with keywords and lower row is for image query without keywords.