Expert Systems With Applications 77 (2017) 20–33

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Featured correspondence topic model for semantic search on social image collections

Nguyen Anh Tu, Kifayat Ullah Khan, Young-Koo Lee*

Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea

ARTICLE INFO

Article history:
Received 20 June 2016
Revised 12 January 2017
Accepted 27 January 2017
Available online 31 January 2017

Keywords:
Image retrieval
Image annotation
Social image tagging
Topic modeling
Probabilistic graphical model

ABSTRACT

Nowadays, due to the rapid growth of digital technologies, huge volumes of image data are created and shared on social media sites. User-provided tags attached to each social image are widely recognized as a bridge to fill the semantic gap between low-level image features and high-level concepts. Hence, a combination of images along with their corresponding tags is useful for intelligent retrieval systems, those are designed to gain high-level understanding from images and facilitate semantic search. However, user-provided tags in practice are usually incomplete and noisy, which may degrade the retrieval performance. To tackle this problem, we present a novel retrieval framework that automatically associates the visual content with textual tags and enables effective image search. To this end, we first propose a probabilistic topic model learned on social images to discover latent topics from the co-occurrence of tags and image features. Moreover, our topic model is built by exploiting the expert knowledge about the correlation between tags with visual contents and the relationship among image features that is formulated in terms of spatial location and color distribution. The discovered topics then help to predict missing tags of an unseen image as well as the ones partially labeled in the database. These predicted tags can greatly facilitate the reliable measure of semantic similarity between the query and database images. Therefore, we further present a scoring scheme to estimate the similarity by fusing textual tags and visual representation. Extensive experiments conducted on three benchmark datasets show that our topic model provides the accurate annotation against the noise and incompleteness of tags. Using our generalized scoring scheme, which is particularly advantageous to many types of queries, the proposed approach also outperforms state-of-the-art approaches in terms of retrieval accuracy.

1. Introduction

With the ever-growing popularity of digital photography, online media sharing sites and social networks (e.g. Flickr, Picasa, and Facebook) have quickly become a powerful part of the Internet with over billions of images uploaded by users. Consequently, the number of stored image data is growing rapidly. Retrieving relevant images from such large collection poses great challenges because there exist various types of information such as text and image feature. A large amount of research effort has been made to design image retrieval systems. Generally, image retrieval research can be categorized into two types of approaches. The first approach is content-based image retrieval (CBIR) (Lew, Sebe, Djeraba, & Jain, 2006; Philbin, Chum, Isard, Sivic, & Zisserman, 2007), which highly depends on visual features like colors or textures to derive a representation of image contents and estimate visual similarity between images. Despite the high efficiency for indexing in the large-scale database, the performance of CBIR is usually limited due to the semantic gap between low-level features and high-level concepts of each image. The second approach is known as text-based image retrieval, which has recently attracted significant focus from researchers in computer vision community (Wu, Jin, & Jain, 2013). This type of approach offers semantic search with the query in the form of natural language. In general, the semantic contents of an image are usually described by manual tags. However, manual tagging is extremely laborious in a huge database. Moreover, the reliability of manual tagging is not guaranteed because tags provided by users can be noisy, incomplete, and sometimes irrelevant.

To address aforementioned issues of manual tagging, there exist extensive research studies to design automatic annotation systems for inferring the tags. One of the most successful approach is to employ statistical generative models (Blei & Jordan, 2003; Carneiro, Chan, Moreno, & Vasconcelos, 2007; Monay & Gatica-Perez, 2004; Wang, Blei, & Li, 2009) to study the correlation between image

* Corresponding author.
E-mail addresses: tunguyen@khu.ac.kr (N.A. Tu), kualizai@khu.ac.kr (K.U. Khan), yklee@khu.ac.kr (Y.-K. Lee).

http://dx.doi.org/10.1016/j.eswa.2017.01.055
0957-4174/

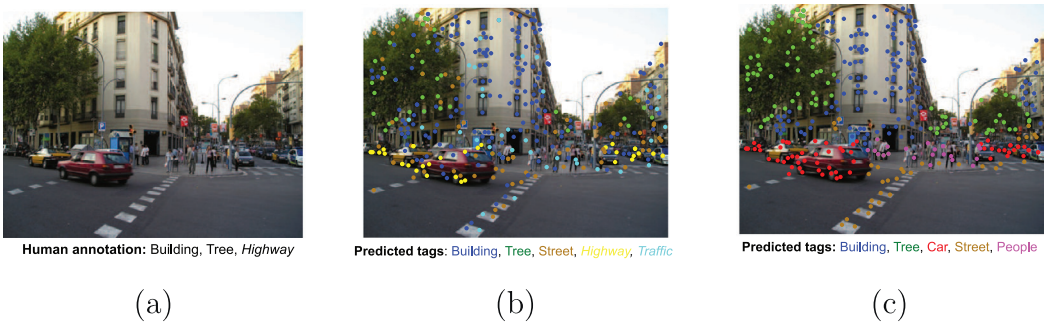

Fig. 1. An example of the resulting image after applied to the CorrLDA and FeaCorrLDA models. The dots represent the visual words extracted by the affine variant detectors. Different colors of dots and tags denote different topics, where each topic is associated with an object appeared in the image. The italic word "highway" indicates a noisy tag. We can observe that by incorporating the relationships between visual words in terms of spatial location and color, the topic extraction of FeaCorrLDA performs much better than that of CorrLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

and text description by using common latent variables. These models generally scale well to database size and the number of tags. Particularly, the Correspondence LDA (CorrLDA) model in Blei and Jordan (2003) and Wang et al. (2009) provides a natural way to learn latent topics from tags and image features. This allows us to encode human knowledge as well as deal with synonyms and homonyms in annotation since each topic reflects a multinomial distribution over the word vocabulary. Moreover, the correspondence topic model can directly tackle the multi-labeling problem, hence it is very efficient to apply in real-world applications. However, methods described in Blei and Jordan (2003) and Wang et al. (2009) focus on employing CorrLDA for image annotation rather than image retrieval as our interest here. In addition, CorrLDA completely ignores useful characteristic of image features such as location and color. As a result of these shortcomings, the effectiveness of accurate topic extraction is reduced. This may lead to the wrong prediction for image annotation through extracted topics as illustrated in Fig. 1(b). On the other hand, the low accuracy of tagging negatively affects the performance of image retrieval when we directly use the output of image annotation (i.e. predicted tags) for indexing and retrieving relevant images. Therefore, semantic image retrieval demands further improvement of image annotation.

In this paper, we propose an image retrieval framework that enables not only the effective image annotation but also an efficient similarity measurement for ranking the retrieved images. First, each image is modeled as a set of local features (e.g. SIFTs Lowe, 2004), and these features are quantized to form a set of visual words. We then propose a probabilistic topic model learned on social images to extract the semantic topics from the co-occurrence of tags and visual words. Our proposed topic model called Featured Correspondence LDA (FeaCorrLDA) extends CorrLDA by incorporating new notions related to the characteristic of visual words. In the proposed model, the correspondence between an image and its textual tags is greatly enforced via the latent topics that are generated from visual features. We further exploit the intuition that the visual words belonging to the same topic are close together in terms of location and color distribution. Then, by formulating simultaneously spatial location and color relationship among visual features, FeaCorrLDA overcomes the limitations of CorrLDA, and therefore improves the topic extraction as shown in Fig. 1(c). As a result, it allows us to effectively complete the missing tags as well as remove the noisy ones.

Since the images are completely annotated by our topic model, the predicted tags then help to achieve our target of retrieving a list of semantically relevant images. In other words, these tags are used for the similarity measurement to rank database images matching a given query in descending order. Accordingly, we present a generalized scoring scheme that handles various types of queries such as keywords, images, and combination of both. This is different from most of the existing semantic search systems (Blei & Jordan, 2003; Guillaumin, Mensink, Verbeek, & Schmid, 2009; Li, Snoek, & Worring, 2009), where they only use keywords to perform searching. In some contexts, the annotations are too general or the users want to find more visually relevant images, hence satisfactory results of image retrieval are hardly obtained with only textual information. This limitation motivates us to fuse both the predicted tags and visual contents for estimating the similarity between the query and database images. In particular, we estimate the textual similarity using tag correspondence measured by the probabilities of predicted tags. At the same time, we estimate the visual similarity based on global image representation, which is formed as a vector converted from the local features. This image representation is computed by using encoding techniques like Fisher Vector (Sánchez, Perronnin, Mensink, & Verbeek, 2013). To boost the retrieval performance, we also integrate color information into encoding scheme. Finally, both textual and visual similarities are combined together to maintain the reliable search.

Our retrieval framework based on correspondence topic model has the following advantages over existing approaches: (1) Due to the nature of the generative model, our model is well-suited to be learned with weakly labeled data by directly exploiting social images available on the Internet. As a result, it is more convenient and applicable to realistic applications with little human labor. (2) Extracting latent topics is an effective way to reduce the semantic gap and guarantee the accurate tag prediction since it allows to enrich more semantics in the scenes. (3) Various prior knowledge about visual cues (e.g., shape, color, and location) can be flexibly included into our model to better address the topic extraction issue. Nevertheless, limitation of topic model is that some parameters (e.g., the number of topics) must be fixed or empirically estimated, and hence it may not generalize well to the newly updated data.

In summary, our major contributions are as following: (1) We develop a generative topic model by encoding the relationship among visual features and exploiting the correlation between tags and image contents. This model enables us to effectively predict the missing tags from an unseen image as well as complete the ones annotated in the challenging databases of social images. (2) We present a scoring scheme based on the fusion of visual and textual information to efficiently compute the similarity between query and database images. Using this scheme, the proposed approach shows very promising results on public datasets. A preliminary version of our work was presented by Tu, Cho, and Lee (2016). In this paper, we modify and further improve the proposed topic model and similarity measuring scheme. We also carry out more empirical studies with the practical setting on real-world datasets.

The remainder of this paper is organized as follows. Section 2 provides an overview of related works. Section 3 introduces our proposed framework, and describes the details of our methods including image preprocessing and problem formulation (Section 3.1), the FeaCorrLDA model (Section 3.2), learning procedure of FeaCorrLDA (Section 3.3), inference of unseen image (Section 3.4), and similarity measurement for image retrieval (Section 3.5). Experimental results on benchmark datasets are conducted and discussed in Section 4. Finally, conclusions are presented in Section 5.

2. Related work

In this section, we briefly review research studies that are closely related to the techniques presented in this paper.

2.1. Image annotation

In recent years, many studies have been carried out in the domain of image annotation. In this line of work, automatic image annotation is considered as a conventional problem, where the objective is to assign a number of semantically related keywords to an unlabeled image. The popular techniques (Jiang, Chang, & Loui, 2006; Qi et al., 2007; Zhou, Cheung, Qiu, & Xue, 2011) treated image annotation as a classification task, where the classifiers were learned from training data to map low-level features into relevant tags. The major disadvantage of these techniques is that their performance is limited to a small vocabulary of tags with well-labeled training data. Different from classification-based method, Carneiro et al. (2007) approached this task as a multi-instance learning problem, where they consider an image as a bag of feature vectors (i.e. instances). They then employed the mixture densities of corresponding semantic labels estimated on the collections of training images to predict tags for the test images. Later, nearest neighbor models with metric learning have been explored in several works (Guillaumin et al., 2009; Li et al., 2009; Wu, Hoi, Zhao, & He, 2011). Li et al. (2009) proposed the voting algorithm under realistic constraints to estimate the correspondence between tags and a given image by counting the occurrence of tag in annotations of visual neighbors. Guillaumin et al. (2009) obtained the state-of-the-art performance of image annotation by using a weighted nearest-neighbor model termed TagProp. Particularly, they proposed a probabilistic framework to compute the probability of using images in visual neighbors based on distance-based weights. To improve the annotation performance, recent works formulated image tagging as multi-label learning problem by incorporating the correlations among labels. A max-margin formulation was built by Hariharan, Zelnik-Manor, Varma, and Vishwanathan (2010) for the correlated predictors, while a discriminative model for multi-class object recognition in Desai, Ramanan, and Fowlkes (2011) was developed in term of the structured prediction. Cao, Zhang, Guo, Liu, and Meng (2015) designed a semantic label dictionary representation under weakly supervised learning that explored the semantic correlation between co-occurrence labels. Besides, Binder, Samek, Müller, and Kawanabe (2013) proposed a random sampling strategy in bag-of-words models to represent the image and improve ranking performance for the image annotation by proposing a kernel multi-task learning method. Although automatic image annotation has obtained the great success in recent developments, there are still existing limitations when dealing with the social database. Since automatic image annotation requires completely annotated images for the training set with clean tags, its performance is easily degraded if training images are partially annotated or contains noisy tags like social images. Therefore, it is difficult to employ automatic image annotation for practical applications having the large-scale database.

Unlike automatic image annotation, tag refinement deals with an image associated with an initial list of tags rather than an unlabeled image. The goal of refining tags is to remove noisy or irrelevant tags from the initial list and enrich it with relevant tags. Wang, Jing, Zhang, and Zhang (2006) formulated tag refinement as the graph ranking problem and applied random work algorithm to re-rank the annotations, where the top ones are considered as the relevant tags. Weinberger, Slaney, and Van Zwol (2008) presented a probabilistic framework to reduce tag ambiguities by finding two tags that appear in different contexts but co-occur in the initial tag list. Liu, Hua, Wang, and Zhang (2010) proposed an optimization framework based on the consistency between "visual similarity" and "semantic similarity" to filter the tags and then enrich the filtered ones by knowledge-based method using WordNet. Lee, De Neve, and Ro (2010) presented the technique to discover a visual folksonomy for an image and refine the noisy tags by using the similarity of images and tag occurrence in the visual folksonomy. Zhu, Yan, and Ma (2010) solved tag refinement problem via a decomposition of user-provided tag matrix, where they took the tagging properties into account from the view of low-rank, error sparsity, content consistency, and tag correlation. The method of Liu, Wu, Zhang, Shao, and Zhuang (2011) treated images and their tags as the compositions of semantic utilities, and then built a semantic unity graph to estimate the relevance between tags and images. In general, above methods work well for tag filtering. However, the tag enrichment is still unsatisfactory, because most of them focus on re-ranking the initial tag list rather than explicitly completing the tags of an image. As a result, their performance on tag-dependent applications is degraded due to the incompleteness of tags. Similar to tag refinement, tag completion directly addresses the issues of manual tags that are incomplete and noisy, but they simultaneously remove the irrelevant tags and predict the missing ones. More recently, a few studies have been made to fit the scope of tag completion as our interest here. Zhuang and Hoi (2011) presented an optimization framework that exploited both semantic and visual contents of social images to discover the correlation between tags and images. Wu et al. (2013) cast tag completion into matrix completion problem by searching for the optimal tag matrix consistent with both the observed tags and visual similarities. Lin, Ding, Hu, Wang, and Ye (2013) simultaneously considered image reconstruction and tag reconstruction to measure the tag relevance scores by using sparse learning techniques. Xia, Feng, Peng, Wu, and Fan (2015) developed the regularized optimization framework that jointly combines three components including the non-negative matrix factorization, holistic visual diversity, and regularization terms.

2.1.1. Topic model for image annotation

Probabilistic topic model has been originated from natural language processing, and widely adapted to image understanding problems. Most studies have addressed the tasks of image classification (Rasiwasia & Vasconcelos, 2013; Wang et al., 2009) and automatic image annotation (Blei & Jordan, 2003; Carneiro et al., 2007; Monay & Gatica-Perez, 2004; Wang et al., 2009). Particularly, for the latter, most works learned the statistical generative model incorporating latent variables (e.g. topics) to formulate a joint distribution between visual features and tags. There have been limited studies using topic models for more challenging tasks like tag refinement and completion. Bundschus et al. (2009) proposed a probabilistic topic model integrated both tag and user information to estimate the posterior probability of the relevant tag for collaborative tagging system. Krestel and Fankhauser (2009) adopted LDA to extract latent topics from resources and then recommend the most relevant tags from these topics. However, visual information (e.g. low-level features) has not been considered in these studies (Bundschus et al., 2009; Krestel & Fankhauser, 2009), since

they only focus on exploiting textual information associated with social images. Without visual information, it is difficult to guarantee the correspondence between tags and the image content. To overcome this limitation, Wang et al. (2014) extended LDA model for tag refinement to force images with similar visual content to have similar topics. Consequently, the tag correspondence is measured as the dot product between topic vectors of image and tag. Similar to other methods for tag refinement, aforementioned studies have not explicitly tackled the problem of missing tags. Unlike existing methods, in this paper, we present the probabilistic correspondence topic model to directly address the specific problem of tag completion.

2.2. Semantic image retrieval

Semantic image retrieval represents the image content by high-level features like keywords. It enables users to specify the query via a textual description that describes the visual concepts of their interest. Therefore, image annotation has been used extensively to facilitate semantic image retrieval (Carneiro et al., 2007; Guillaumin et al., 2009; Li et al., 2009; Wu et al., 2013). In particular, retrieval systems based on automatic image annotation (Carneiro et al., 2007; Guillaumin et al., 2009; Li et al., 2009) learn semantic concept models from the training images that are completely labeled with relevant tags. Database images, which are annotated with semantic labels by employing concept models, can be retrieved by semantic keywords similar to text document retrieval. However, to maintain high retrieval accuracy, these annotation-based approaches have the same limitation as automatic image annotation that essentially needs a large number of training images which are manually labeled with clean tags. Consequently, most of the recent studies draw more attention to tag completion (Wu et al., 2013; Xia et al., 2015; Zhuang & Hoi, 2011) to directly train models from the partially annotated images (e.g. social images with noisy and incomplete tags), and efficiently complete the tags of database images as well as new images. On the other hand, we aim to develop the retrieval framework enabling various types of queries. This is closely related to cross-media retrieval (Ding, Guo, & Zhou, 2014; Song, Yang, Yang, Huang, & Shen, 2013; Yang, Xu, Nie, Luo, & Zhuang, 2009; Yang, Zhuang, Wu, & Pan, 2008; Zhuang, Yang, & Wu, 2008), which also allows users to query different media types. For example, Zhuang et al. (2008) built a heterogeneous graph of multimedia for multimedia retrieval, where nodes represent media objects of different modalities and weighted edges represent the correlation among nodes. The main disadvantage of traditional techniques in cross-media retrieval is that they are not scalable for the large database. To improve the scalability, recent studies has been attracting to developing indexing algorithm for supporting efficient multimedia retrieval. Song et al. (2013) proposed the hashing model exploring inter-view and intra-view consistency to learn linear hash functions for mapping features in different views into a common Hamming space. The unified hash codes from different modalities were learned in Ding et al. (2014) by collective matrix factorization. Different from these cross-media retrieval approaches, which mainly focus on indexing stage, our work is based on the annotation task along with the capability of tag completion to deploy an efficient scheme of similarity measurement. Moreover, the existing annotation-based and completion-based approaches can only perform searching with text queries, while our approach allows us to deal with different query types (e.g., text and image). The theoretical comparison between the proposed research with highlight approaches in semantic image retrieval is further summarized in Table 1.

Generally, our work inherits the advantages of topic model which possesses the great ability in encoding expert knowledge via topics and capturing the multi-label property of image annotation. Here, semantic topics follow the multinomial distributions over the textual and visual vocabularies, those are highly correlated and enable the great degree of correspondence between features and tags (labels). Moving beyond image annotation, we further show that the correspondence topic model is also suitable for tag completion, which has not been studied in previous topic modeling approaches. This is achieved by addressing the issue of topic extraction, where we investigate how the performance of annotation and completion tasks is affected by the quality of extracting semantic latent topics. Moreover, motivated by the success of cross-media retrieval in handling various query types, we also study how integrating visual and textual information into scoring scheme can help to provide a robust retrieval system.

3. Proposed methodology

We now present our proposed methodology in this section and illustrate its framework in Fig. 2. Our proposed framework is composed of two main components i.e., offline and online processes. The offline process extracts the topics from each social image to estimate the tag correspondence and complete manually annotated tags. During this stage, the model parameters (i.e., textual and visual topic distributions) of FeaCorrLDA model are learned from the training images. Specifically, each topic is discovered and modeled as the distributions over visual features and tags by capturing the co-occurrence patterns of image contents and corresponding texts. Otherwise, the FeaCorrLDA model performs topic extraction based on the intuition that, within an image, visual features (e.g., the parts of an object) close together in terms of location and color distribution tend to be assigned to the same topic. Hence, the learned parameters (via topics) implicitly reflect the knowledge about the correlation between tags and image contents as well as the relationship among visual features. These parameters also represent how consistent the discovered topics are with the semantic concepts (from tags) as shown in Fig. 2. Subsequently, the online process aims to retrieve the ranked list of relevant images with respect to the similarity scores between query and database images, which is computed using the visual content and predicted tags. In this stage, FeaCorrLDA adopts learned parameters to accelerate the search speed. Note that the query can be image, keywords, or combination of both. If the query only contains keywords, it can be directly measured against the completed tags of database images without applying to topic model. The different parts of the framework will be described in more detail in the following subsections.

3.1. Feature extraction and problem formulation

3.1.1. Feature extraction

In this part, we present main steps of feature extraction to preprocess an image as illustrated in Fig. 3. We first find salient interest points in the image by using an affine invariant detector (Lowe, 2004). The image patches around each interest point are then normalized and used to compute two types of descriptors including SIFT descriptor for shape representation and hue descriptor (Van De Weijer & Schmid, 2006) for color representation. Let $\mathbf{b} = \{b_i\}_{i=1}^{N}$ be a set of SIFT descriptors, and $\mathbf{c} = \{c_i\}_{i=1}^{N}$ be a set of hue descriptors in the given image. Each 128-D SIFT descriptor is quantized (Philbin et al., 2007) to a visual word v via visual vocabulary or codebook learned by k-means clustering. Moreover, each hue descriptor is combined with location (l_x, l_y) of its patch to form a regional feature f.

3.1.2. Problem formulation

A social image $\mathbf{I_d}$ from the training data $\mathbb{B} = \{\mathbf{I_d}\}_{d=1}^{D}$ is represented by two types of entities: text words (or tags) $\mathbf{w_d} = \{w_{di}\}_{i=1}^{M_d}$

Table 1
Theoretical comparison of semantic image retrieval approaches.

Method	Main task	Training tags	Query type	Key technique
Carneiro (2007)	Annotation	Clean	Text	Gaussian mixture model
Guillaumin (2009)	Annotation	Clean	Text	Weighted nearest neighbor
Li (2009)	Annotation	Clean	Text	Neighbor voting
Zhuang (2011)	Annotation Completion	Noisy Incomplete	Text	Stochastic coordinate descent
Wu (2013)	Annotation Completion	Noisy Incomplete	Text	Subgradient descent
Song (2013)	Indexing	Noisy Incomplete	Text Image	Inter-media hashing
Ding (2014)	Indexing	Noisy Incomplete	Text Image	Collective matrix factorization
Xia (2015)	Completion	Noisy Incomplete	Text	Non-negative matrix factorization
Proposed	Annotation Completion	Noisy Incomplete	Text Image	Correspondence topic model

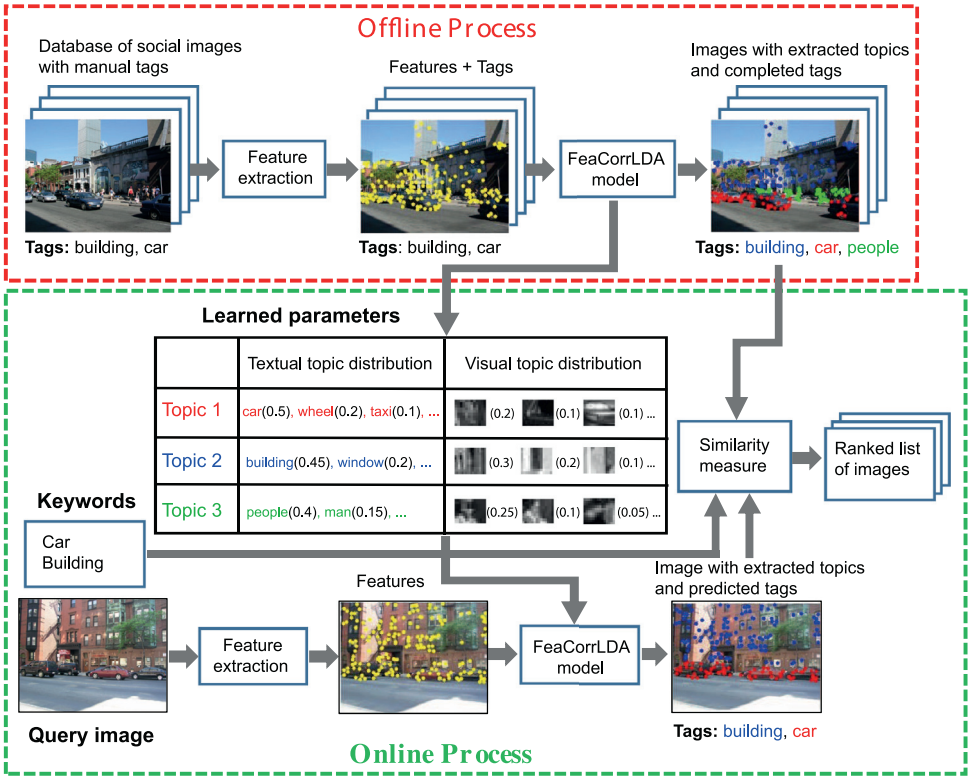

Fig. 2. Overview of proposed framework. Our method consists of offline and online processes. The offline process learns latent topics from social images and then completes the tags of database images. The online process retrieves the relevant images by measuring the similarity between the query (e.g. keywords, image) and the completely annotated images in the database. Yellow dots denote features before applied to FeaCorrLDA. Red, blue, and green dots denote features assigned to topic 1, topic 2, and topic 3, respectively, after applied to FeaCorrLDA. Learned parameters consist of textual and visual distributions of discovered topics. Each topic shows the examples of top-ranked tags and top-ranked features, which are sorted by their probability in textual and visual distributions, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and local features consisting of a bag-of-visual words $\mathbf{v_d} = \{v_{di}\}_{i=1}^{N_d}$ and a set of regional features $\mathbf{f_d} = \{f_{di}\}_{i=1}^{N_d}$. Given the training data $\mathbb{B}$ and a predefined vocabulary of text word $\mathbb{C} = \{word_i\}_{i=1}^{W}$, our objective is to learn the model for semantic search that automatically completes the tags of social images as well as annotates the unseen images (e.g. query, database images). The problems are formalized as following:

- Tagging image: How to compute the probability $P(w|\mathbf{I}, \mathbb{B})$ conditioned on a given image $\mathbf{I}$ and training data $\mathbb{B}$, where $w = 1 \ldots W$ corresponds to the index of text word in vocabulary. The most relevant tags are then sorted in term of the computed probability. Otherwise, how to guarantee that the probabilities of noisy tags are small and these tags should be ranked lower.
- Similarity measurement: How to efficiently estimate the semantic similarity $S(q, d)$ between query and database images.

3.2. Featured correspondence LDA

The CorrLDA model (Blei & Jordan, 2003; Wang et al., 2009) (as shown in Fig. 4(a)) is the generative model for multiple entity data like the social images, where each one is formed as the combination of visual words and text words. However, it is difficult to obtain the desired result for tag correspondence, if we naively apply conventional CorrLDA. The reason is that this model com-

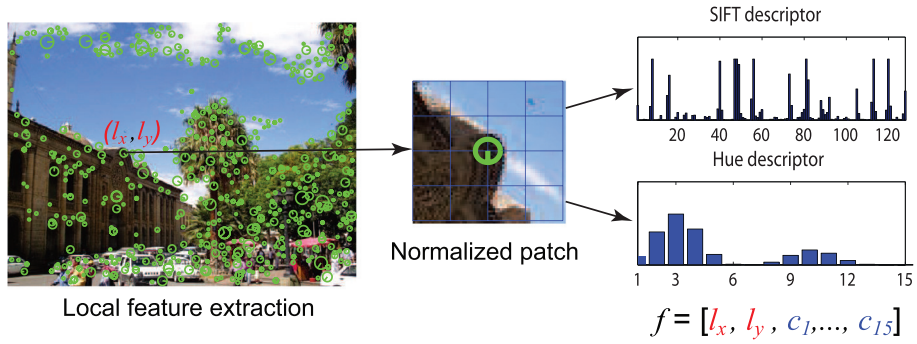

Fig. 3. The process of extracting region of interest (patch) to compute SIFT and hue descriptors. The regional feature f is formed by concatenating hue vector with a location of patch represented by coordinate (l_x, l_y).

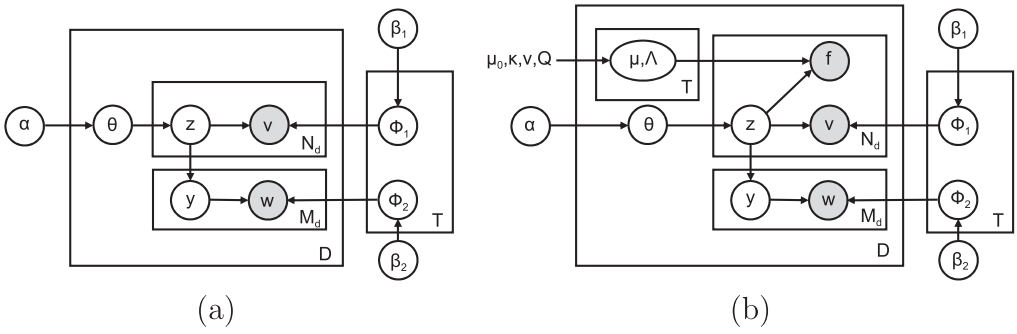

Fig. 4. Probabilistic topic model: (a) correspondence LDA (CorrLDA) and (b) featured correspondence LDA (FeaCorrLDA).

pletely ignores important characteristics of image features (e.g. location, color), and hence reduce the effectiveness of topic extraction. Therefore, in this study, we extend the CorrLDA model by incorporating new notions related to the characteristic of visual words that helps the proposed model to be more suitable for tackling specific vision problems.

Our proposed topic model is illustrated by the graphical model shown in Fig. 4(b). It represents a collection of D images, and each image $\mathbf{I_d}$ consists of N_d visual words and M_d text words. We use latent variables (i.e., z_{di}) to characterize the topics, where topic z indicates a semantically related term for each visual feature. In addition, we factor the image into a combination of T topics. Each topic is modeled as two types of distributions over the visual vocabulary of size V and over the textual vocabulary of size W. The textual topic (denoted by latent variable y) is a counterpart of topic z. Thus, the FeaCorrLDA model directly uses the latent topic of visual words for generating the text words. According to the graphical model, v_{di} and w_{dj} are the observed variables. We further introduce new observed variable f_{di} to represent the characteristic of i^{th} image patch which is measured by the regional feature vector as described in Section 3.1. Formally, the generative process of our FeaCorrLDA model for image corpus is as follows:

1. For each topic t:
   (a) Draw an appearance distribution $\phi_{1,t} \sim Dir(\beta_1)$
   (b) Draw an textual distribution $\phi_{2,t} \sim Dir(\beta_2)$
2. For each image $\mathbf{I_d}(d = 1, \ldots, D)$:
   (a) Draw topic proportion $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each topic $t(t = 1, \ldots, T)$, draw a regional feature distribution: $\{\mu_{td}, \Lambda_{td}\} \sim NW(\mu_0, \kappa, \nu, Q)$
3. For each visual word v_{di} where $i \in 1, 2, \ldots, N_d$:
   (a) Draw topic $z_{di} \sim \text{Multi}(\theta_d)$
   (b) Draw visual word $v_{di} \sim \text{Multi}(\phi_{1,z_{di}})$
   (c) Draw regional vector $f_{di} \sim N(\mu_{dz_{di}}, \Lambda^{-1}_{dz_{di}})$
4. For each text word w_{dj} where $j \in 1, 2, \ldots, M_d$:
   (a) Draw topic $y_{dj} \sim \text{Unif}(z_1, \ldots, z_{Nd})$
   (b) Draw text word $w_{dj} \sim \text{Multi}(\phi_{2,y_{dj}})$

Here, Dir, Multi, N, and NW denote Dirichlet, Multinomial, Normal, and Normal-Wishart distributions, respectively. The priors including Multi and NW are chosen to conjugate to Dir and N for the word and location distributions, and hence, they simplify computation and guarantee efficient inference. In this model, we employ two types of entities: visual entity (i.e. visual words and regional features) and textual entity (i.e. tags). Visual words and text words are discrete random variables, and hence follow Multi distribution. Regional features are real-valued vectors, and so modeled as Normal distribution.

Our FeaCorrLDA model allows us to gain insight into prior knowledge from human expertise, which is systematically exploited as Bayesian priors about visual appearance and the relationship between entities. In particular, as described in generative process, each topic t in image d is represented by a probability distribution $p(\mu_{td}, \Lambda_{td})$ of regional features and a probability distribution Φ_t of visual words. The distribution of regional features consisting of color and location $p(\mu_{td}, \Lambda_{td})$ is not shared among images, whereas the distribution of visual words $\Phi_{1,t}$ are shared across images. The reason is that the shape appearance of an object (e.g. "car", "tree", "people") captured by the $\Phi_{1,t}$ distribution is similar in all images. In contrast, the location and color distributions of an object in a specific image can be assumed to be independent of the color and location in other images. Moreover, the relationship among visual words in an image is encoded via the parameters μ_{td}, Λ_{td}, where the visual words belong to the topic t along with their regional features that are close to the expected value of distribution $p(\mu_{td}, \Lambda_{td})$. Otherwise, the textual 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_2). Therefore, the correlation of visual and text words is highly enforced using this model.

3.3. Parameter estimation in FeaCorrLDA

In this subsection, we describe a learning method for the proposed model. Let $\Pi = \{\alpha, \beta_1, \beta_2, \mu_0, \kappa, \nu, Q, \}$ be the set of hyper parameters. Given a training data $\mathbb{B}$ of D images, the model parameters including $\Phi_1 \in \mathbb{R}^{V\times T}$ of visual topic distribution and $\Phi_2 \in \mathbb{R}^{W\times T}$ of textual topic distribution can be estimated by maximizing the following log likelihood function.

$$L(\Phi_1, \Phi_2) = \sum_{d=1}^{D} \log\left(p(\mathbf{v_d}, \mathbf{f_d}, \mathbf{w_d}, \mathbf{z_d}, \mathbf{y_d}|\Phi_1, \Phi_2, \Pi)\right) \tag{1}$$

However, it is intractable to directly estimate from the distribution in Eq. (1). Instead, we use Monte Carlo EM algorithms (Andrieu, De Freitas, Doucet, & Jordan, 2003) as summarized in Algorithm 1 for approximate estimation. The parameters are then

Algorithm 1 Parameter estimation of FeaCorrLDA.

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

Output: The estimated parameters Φ_1 and Φ_2

1. **Initialization**. Initialize set of parameters $\left\{\Phi_1^{(0)}, \Phi_2^{(0)}\right\}$
2. **For each** $k = 1, \ldots, K$ **do**:
   (a) Given $\Phi_1^{(k-1)}$, for each image $\mathbf{I_d}$, sample latent variables with N Gibbs steps from the posterior distribution of visual topic $p(\mathbf{z_d}|\mathbf{v_d}, \mathbf{f_d}, \mathbf{y_d}, \Pi)$ using Eq. (2).
   (b) Given $\Phi_2^{(k-1)}$, for each image $\mathbf{I_d}$, sample latent variables with M Gibbs steps from the posterior distribution of textual topic $p(\mathbf{y_d}|\mathbf{w_d}, \mathbf{z_d}, \Pi)$ using Eq. (3).
   (c) Compute $\left\{\Phi_1^{(k)}, \Phi_2^{(k)}\right\}$ using as Eqs. (5) and (6).
3. **End**

estimated by examining the posterior samples. Here, we employ the collapsed Gibbs sampling algorithm (Griffiths & Steyvers, 2004) for sampling of latent variables z and y for visual word v and text word w, as in the following equations:

$$\begin{aligned} p(z_{di}=t|\mathbf{v_d}, \mathbf{f_d}, \mathbf{z_{-di}}, \mathbf{y_d}, \Pi) &\propto p(v_{di}|\mathbf{v_{-di}}, \mathbf{z_d}, \Pi) \times p(f_{di}|\mathbf{f_{-di}}, \mathbf{z_d}, \Pi) \\ &\times p(z_{di}=t|\mathbf{z_{-di}}, \Pi) \times p(\mathbf{y_d}|z_{di}=t, \Pi) \\ &\propto \frac{n_{vt,-di}^{VT} + \beta_1}{\sum_{v'} n_{v't,-di}^{VT} + V\beta_1} \times t\left(\mu_{0,td,-di}, \frac{Q_{td,-di}(\kappa_{td,-di}+1)}{\kappa_{td,-di}(\nu_{td,-di}-q+1)}\right) \\ &\times \frac{n_{td,-di}^{TD} + \alpha}{\sum_{t'} n_{t'd,-di}^{TD} + T\alpha} \times \prod_{j=1}^{M_d} \frac{n_{td,-di}^{TD} + \mathbb{I}(y_{dj}=t)}{\sum_{t'} n_{t'd,-di}^{TD} + \mathbb{I}(y_{dj}=t)} \end{aligned} \tag{2}$$

$$\begin{aligned} p(y_{dj}=t|\mathbf{w_d}, \mathbf{z}, \mathbf{y_{-di}}, \Pi) &\propto p(w_{dj}|\mathbf{w_{-dj}}, \mathbf{y_d}, \Pi) \times p(y_{dj}=t|\mathbf{z}, \Pi) \\ &\propto \frac{n_{wt,-dj}^{WT} + \beta_2}{\sum_{w'} n_{w't,-di}^{WT} + W\beta_2} \times \frac{n_{td}^{TD}}{N_v} \end{aligned} \tag{3}$$

where the subscript *-di* indicates whole variables excluding the i^{th} variable in image d; 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_{wt}^{WT} is the number of times text word w is assigned to topic t; $\mathbb{I}$ is the indicator function; q is the dimensionality of regional features. The second term of Eq. (2) denotes the t-student distribution with a set of parameters μ_0, ν and Q (see Murphy, 2012, chapter 4.6.3 for detailed computation of these parameters).

After some sampling iterations, parameters Φ_1, Φ_2 are estimated until convergence by using posterior samples of latent variables. The posterior of the topic-visual word multinomial is computed as below:

$$p(\Phi_{1,t}|\mathbf{v}, \mathbf{z}) = \mathrm{Dir}\left\{\beta_1 + n_{vt}^{VT}\right\} \tag{4}$$

where $\mathbf{v} = \{\mathbf{v_d}\}_{d=1}^{D}$, $\mathbf{z} = \{\mathbf{z_d}\}_{d=1}^{D}$. Thus, Φ_1 can be estimated as the posterior mean of $p(\Phi_{1,t}|\mathbf{v}, \mathbf{z})$, which is simply the normalized Dirichlet parameters, as follows:

$$\Phi_{1,t} = \frac{n_{vt}^{VT} + \beta_1}{\sum_{v'} n_{v't}^{VT} + V\beta_1} \tag{5}$$

Similarly, we can estimate Φ_2 of the textual topic distributions as follows:

$$\Phi_{2,t} = \frac{n_{wt}^{WT} + \beta_2}{\sum_{w'} n_{w't}^{WT} + W\beta_2} \tag{6}$$

3.3.1. Remarks on tag completion of training images

Based on Eqs. (2) and (3), we can see how topic assignment affects the task of tag completion by generating the missing tags and removing the noisy ones simultaneously. In particular, the first term of Eq. (2) is the probability of visual word v assigned to topic t, while the second and third terms are the probability of topic t for image d with respect to regional features and visual words. Hence, these factors force visual words that co-occur in the same image and correlate each other through regional features to be assigned to the same topic. Otherwise, the last term of Eq. (2) is the assignment probability of text word given the topic assignment of visual word. As a result, manual tags that frequently appear with topic t in the same image tend to be assigned to this topic. Thus, by using the extracted topics in each image, we can effectively predict or generate the missing tags. Eq. (3) further measures the probability of a text word assigned to topic t for image d. While the first term is the probability of text word w assigned to topic t in the entire collection, the second term indicates the proportion of topic t in image d. Based on the fact that noisy words caused by personalized annotation rarely appear in the image collection, it will significantly decrease the probability of assigning noisy word w to topic t. On the other hand, if topic t dominates in image d, it will increase the probability of assigning text word to topic t. Therefore, using the assignment probabilities, we can directly deal with the tag completion problem, since the noisy words will be ranked lower whereas the semantically related words will be added or ranked higher via the extracted topics.

3.4. Inference of unseen image and tag prediction

In inference procedure, our objective is to infer the latent variables (i.e. z and y) of unseen image $\mathbf{I_{d'}}$. Here, the inference procedure is similar to parameter estimation except that we replace the first terms of Eqs. (2) and (3) by learned parameters $\Phi_{1,t}$ and $\Phi_{2,t}$, as shown in Eqs. (7) and (8). Consequently, all terms of these sampling equations are only related to a single image. Since individual unseen images are considered independently, our proposed model is applicable to parallel processing. This is practical in real-world scenario where we need to cope with the very large number of social images.

$$\begin{aligned} p(z_{d'i}=t|\mathbf{v_{d'}}, \mathbf{f_{d'}}, \mathbf{z_{-d'i}}, \mathbf{y_{d'}}, \Pi) &\propto \Phi_{1,t} \\ &\times t\left(\mu_{0,td',-d'i}, \frac{Q_{td',-d'i}(\kappa_{td',-d'i}+1)}{\kappa_{td',-d'i}(\nu_{td',-d'i}-q+1)}\right) \\ &\times \frac{n_{td',-d'i}^{TD'} + \alpha}{\sum_{t'} n_{t'd',-d'i}^{TD'} + T\alpha} \times \prod_{j=1}^{M_{d'}} \frac{n_{td',-d'i}^{TD'} + \mathbb{I}(y_{d'j}=t)}{\sum_{t'} n_{t'd',-d'i}^{TD'} + \mathbb{I}(y_{d'j}=t)} \end{aligned} \tag{7}$$

$$p(y_{d'j}=t|\mathbf{w_{d'}}, \mathbf{z}, \mathbf{y_{-d'i}}, \Pi) \propto \Phi_{2,t} \times \frac{n_{td'}^{TD'}}{N_v} \tag{8}$$

Given the FeaCorrLDA model with estimated model parameters, we further propose the algorithm for estimating tag correspondence as shown in Algorithm 2. Particularly, the correspondence

Algorithm 2 Tag correspondence of FeaCorrLDA.

Input: Unseen image formed as a bag of visual words, regional features, and text words $\{\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \mathbf{w}_{\mathbf{d}'}\}$ with learned parameters Φ_1 and Φ_2

Output: The probability of tag correspondence $P(w|\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \Phi_1, \Phi_2)$

1. Sample latent variables with N Gibbs steps from the posterior distribution of visual topic $p(\mathbf{z}_{\mathbf{d}'}|\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \mathbf{y}_{\mathbf{d}'}, \Pi)$ using Eq. (7).
2. Sample latent variables with M Gibbs steps from the posterior distribution of textual topic $p(\mathbf{y}_{\mathbf{d}'}|\mathbf{w}_{\mathbf{d}'}, \mathbf{z}_{\mathbf{d}'}, \Pi)$ using Eq. (8).
3. Compute $P(w|\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \Phi_1, \Phi_2)$ using as Eq. (9).
4. **End**

of a tag w for each image $\mathbf{I}_{\mathbf{d}'}$ is formulated as the probability conditioned on the set of regional features $\mathbf{f}_{\mathbf{d}'}$, visual words $\mathbf{v}_{\mathbf{d}'}$, and estimated parameters from training data $\mathbb{B}$. It can be computed as follows:

$$P(w|\mathbf{I}_{\mathbf{d}'}, \mathbb{B}) = P(w|\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \Phi_1, \Phi_2) = \sum_t P(w|t, \Phi_2)P(t|\mathbf{v}_{\mathbf{d}'}, \mathbf{f}_{\mathbf{d}'}, \Phi_1, \Phi_2) = \sum_t \Phi_{2,wt}\theta_{td'} \tag{9}$$

Hence, tag prediction for image $\mathbf{I}_{\mathbf{d}'}$ can be performed by a dot product between the w^{th} row of matrix of tag-topic distribution Φ_2 and the d^{th} column of matrix of topic-image proportion θ. Here, topic-image proportion $\theta_{d'}$ for each image $\mathbf{I}_{\mathbf{d}'}$ is estimated by examining posterior samples of $p(\theta_{d'}|\mathbf{z}_{\mathbf{d}'}) = Dir\{\alpha + n_{td'}^{TD'}\}$, which associates with the third term of Eq. (7). The most relevant tags are then ranked based on the computed probability of tag correspondence. Furthermore, Eq. (9) is derived from the sampling equations, which are the modification of Eqs. (2) and (3). Consequently, as our observation on the sampling equations in Section 3.3, this probability of tag correspondence also enables us to deal with tag completion problem for unseen images.

3.5. Similarity measure using textual and visual information for semantic image retrieval

Given the predicted or completed tags, previous text-based approaches only use tag information for retrieving images. This may lead to unsatisfactory results if the annotation is incorrect or too general. For example, with a query image of "dog", we can retrieve the list of images annotated by a general tag "animal" with very low visual similarity, where the results can be the images of "cat", "tiger", or "bear". This limitation motivates us to fuse the visual and textual information to achieve more reliable search. In this work, we represent the visual content information by Fisher Vector (Sánchez et al., 2013), which uses Gaussian Mixture Model (GMM) to aggregate all local features of an image into a global vector representation. This encoding technique has shown the superior performance for retrieval and classification tasks.

As described in Section 3.1, an image contains two types of local features including shape-based SIFT descriptors and color-based hue descriptors. It has been shown in literature (Van De Sande, Gevers, & Snoek, 2010) that the combination of shape and color may significantly improve the performance of visual recognition. This motivates us to employ the combined descriptor called hue-SIFT descriptor (Van De Weijer & Schmid, 2006), which concatenates SIFT descriptor and hue descriptor together according to $x_i = [b_i, c_i]$ where $i = 1, \ldots, N$. The computation of Fisher Vector u for representing visual information is given as following:

$$\Psi_k^{(1)}(\mathbf{x}) = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} \gamma_k(x_i)\left(\frac{x_i - \eta_k}{\sigma_k}\right) \tag{10}$$

$$\Psi_k^{(2)}(\mathbf{x}) = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} \gamma_k(x_i)\left(\frac{(x_i - \eta_k)^2}{\sigma_k^2} - 1\right) \tag{11}$$

$$u = \left[\Psi_1^{(1)}(\mathbf{x}), \Psi_1^{(2)}(\mathbf{x}), \ldots, \Psi_K^{(1)}(\mathbf{x}), \Psi_K^{(2)}(\mathbf{x})\right] \tag{12}$$

Here, $\{\pi_k, \eta_k, \sigma_k^2\}_{k=1}^K$ are the mixture weights, means, and diagonal covariances of the GMM, which is constructed during training process. $\Psi_k^{(1)}, \Psi_k^{(2)}$ are the encoding functions of the k^{th} Gaussian for the first and second order statistics of local features, respectively. The soft-assignment $\gamma_k(x_i)$ of the ith feature to the kth Gaussian is computed as the responsibility of the GMM component k:

$$\gamma_k(x_i) = \frac{\pi_k exp\left(-0.5(x_i - \eta_k)^T \sigma_k^{-2}(x_i - \eta_k)\right)}{\sum_j \pi_j exp\left(-0.5(x_i - \eta_j)^T \sigma_j^{-2}(x_i - \eta_j)\right)} \tag{13}$$

Given a query image $\mathbf{I}_{\mathbf{q}}$ and database image $\mathbf{I}_{\mathbf{d}}$, their visual representations are u_q and u_d. Let r_q and r_d be two W-dimensional vectors representing for textual information of images $\mathbf{I}_{\mathbf{q}}$ and $\mathbf{I}_{\mathbf{d}}$, respectively, where the w^{th} element of these vectors correspond to the probability of predicted tag w computed by Eq. (9). Then, the similarity score between two images is defined as follows:

$$S(q, d) = (1 - \rho)u_q u_d + \rho r_q r_d \tag{14}$$

One can observe that the first term $u_q u_d$ corresponds to visual similarity, while the second term $r_q r_d$ corresponds to textual similarity. The parameter $\rho \in [0, 1]$ controls the weight of textual similarity in the above score. The influence of setting this value on the retrieval accuracy will be investigated further in our experimental section. 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 the scoring stage by setting $\rho = 1$. In this case, $r_q \in \{0, 1\}^W$ where its i^{th} element is set to 1 if the i^{th} tag appear in the query. Therefore, by changing parameter ρ, our generalized scoring scheme based on the fusion of visual contents and textual tags can handle different types of queries including keywords, images and combination of both. It should be noted that if the query is an image without keywords, the textual part of the FeaCorrLDA model is excluded and the topics are purely extracted from the visual features.

4. Experiments

This section presents the experimental evaluation of our proposed approach. We first describe the datasets used in our experiments. We then present results for topic learning, image annotation, and image retrieval.

4.1. Datasets and experimental setting

4.1.1. Datasets

In this study, we use three benchmark datasets, whose statistics are summarized in Table 2. The description of each dataset is as following:

- Labelme (Russell, Torralba, Murphy, & Freeman, 2008): 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.

Table 2
Statistics of benchmark datasets.

	Labelme	IAPR TC12	NUS-WIDE
Number of images	2920	19,805	237,131
Vocabulary size of tags (W)	490	291	1000
Tags per image (mean/max)	11/48	5.7/23	6.5/131

- IAPR TC12 (Guillaumin et al., 2009): it consists of 19,805 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.
- NUS-WIDE (Chua et al., 2009): This is a large real-world web dataset consisting of over 269,000 images and associated tags downloaded from Flickr, with more than 5000 unique tags. The maximum number of tags per image is 130. Similar to Chua et al. (2009), we keep the top 1000 most popular tags as its vocabulary to reduce noisy tags, which slightly reduce the size of database to 237,131 images.

4.1.2. Experimental setup

We use Difference-of-Gaussian (DoG) (Lowe, 2004) to find the salient points in the image. To compute Hue descriptor of each image patch, we set the number of bins for hue histogram to 15. Then, the dimension of HueSIFT descriptor is $dim_{HS} = 128 + 15 =$ 143-D. Consequently, the dimension of Fisher Vector is $dim_{FV} = 2 \times K \times dim_{HS} =$ 18304-D, where $K = 64$ is the number of Gaussian components. This dimension is high and not convenient for memory storage, therefore we further reduce it to 1024-D using PCA. For all datasets, the vocabulary size V of visual words is set to 2000. We then design the following experiments.

Topic learning: In our experiments, the hyper parameters of FeaCorrLDA and CorrLDA models are set as following: $\alpha = 0.2, \beta_1 = 0.01, \beta_2 = 0.1$. Using these values, we evaluate and compare the learning ability of two models through the log-likelihood, since it reflects the fitting of topic model with the training data. The higher score of the log-likelihood is better. As mentioned in Section 3.2, we assume that latent topics are generated from visual features. Then, the marginal likelihood of visual words for both models $P(\mathbf{v}|\mathbf{z})$ can be computed by integrating out latent variables as follows:

$$P(\mathbf{v}|\mathbf{z}) = \left(\frac{\Gamma(V\beta_1)}{\Gamma(\beta_1)^V}\right)^T \times \prod_{t=1}^{T} \frac{\prod_v (n_{vt}^{VT} + \beta_1)}{\Gamma\left(\sum_{v'} n_{v't}^{VT} + V\beta_1\right)} \tag{15}$$

Eq. (15) implies that the learning performance is strongly affected by the number of topics T. Based on that, during the training process, it is essential to estimate the ideal number which maximizes the likelihood. In this case, we run the algorithm a number of times under different topic number. Note that this is not a perfect way, since it unnaturally handles a flexible number of topics when the data is dynamically updated. However, to some extent, we can still maintain the well-learned model from enough training data.

Image annotation and semantic image retrieval: The experimental results of image annotation are measured using average precision and average recall, which are denoted as *AP@N* and *AR@N*, respectively, for the top N tags annotated to a test image. The precision and recall of each test image are computed based on the relevant tags and the ground-truth. Then, we average both over all test images. Furthermore, we used mean average precision (mAP) to quantitatively evaluate the retrieval performance of 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.

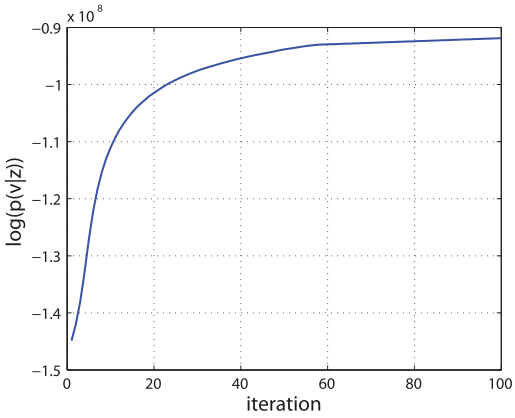

Fig. 5. Convergence of the FeaCorrLDA model. Log-likelihood stabilizes after approximately 60 iterations.

To investigate the improvement of performance using proposed method, we compare with the following state-of-the-art methods: CorrLDA (Blei & Jordan, 2003; Wang et al., 2009), Tag Relevance by Neighborhood Voting (TagNV) (Li et al., 2009), Tag Propagation (TagProp) (Guillaumin et al., 2009), Tag Matrix Completion (TMC) (Wu et al., 2013), and Linear Sparse Reconstruction (LSR) (Lin et al., 2013).

4.2. Topic learning

We first evaluate the performance of topic learning by running Gibbs Sampling during offline process on the training datasets. To estimate the convergence of FeaCorrLDA, we use the training set of IAPR TC12 with 100 topics and run the model for 100 iterations. As shown in Fig. 5, our proposed model converges after approximately 60 iterations. The results of convergence are almost similar when we run on other datasets. We further compare the learning performance of FeaCorrLDA with CorrLDA as shown in Fig. 6, where we measure the log-likelihood of both models over three training datasets by varying the number of topics T from 20 to 300. We can see that log-likelihood of our model is significantly higher than CorrLDA, which shows that our model is more well-fitted to training data and its learning ability is better. Moreover, as the number of topics increase, the log-likelihood of both models increases too. It shows that higher number of topics may improve the performance of topic learning. However, if T is too high, it will result in overfitting and degrade the learning performance. In addition, it requires more running time with the higher number of topics. For our remaining experiments, we empirically select $T = 120$, 140, and 180 as optimal value for IAPR TC12, LabelMe, and NUS-WIDE, respectively, because the log-likelihood over each dataset is not significantly increased after these values.

We further illustrate the extracted topics by using our model as shown in Table 3. In particular, we demonstrate example topics derived from the training set of NUS-WIDE dataset, where each topic lists the top-ranked tags. Accordingly, Table 3 reflects the semantic topic distribution of FeaCorrLDA. We can observe that the topics consistently show the semantic concepts with respect to their tags. For example, topic 3 is related to “pet”; topic 96 is related to “color”; and topic 57 is related to “winter season”. Therefore, it shows the distinct advantage of using topic model to reduce the semantic gap for image annotation. Based on the scene described by extracted topics of an image, more relevant words with high probability can be annotated , whereas the noisy words can be removed or ranked lower.

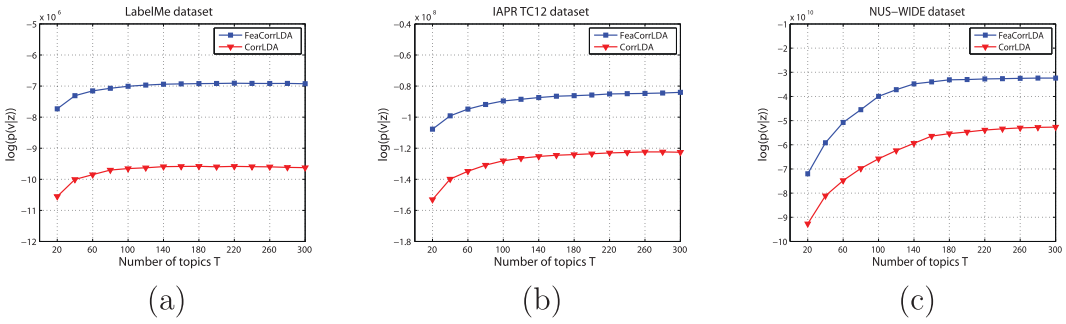

Fig. 6. Likelihood comparison between FeaCorrLDA and CorrLDA w.r.t the number of topics on: (a) LabelMe dataset, (b) IAPR TC12 dataset and (c) NUS-WIDE dataset. The higher score of log-likelihood is, the better topic model fits.

Table 3
Example of extracted topics from NUS-WIDE dataset.

Topic 3	Topic 96	Topic 57	Topic 25	Topic 88	Topic 9	Topic 73	Topic 52
Dog	Yellow	Winter	Architecture	People	Tree	Rain	Airplane
Pet	Orange	Snow	Building	Men	Green	Weather	Flight
Cat	Colors	Ice	Tower	Women	Nature	Cloud	Aircraft
Animal	Gold	Cold	Window	Friend	Branch	Wet	Plane
Puppy	Red	Frozen	Glass	Adult	Leaves	Umbrella	Sky
Friend	Blue	Skiing	Structure	Boy	Forest	Storm	Airport
Cute	Green	Frost	Geometry	Portrait	Mountain	Summer	Fly
Kitty	Light	Glacier	Pattern	Girl	Fields	Sky	Aviation
Squirrel	Brown	Mountain	Lines	Actor	Meadow	Wind	Jet
Adorable	Pink	Iceland	Steel	Group	Landscape	Environment	Wings

4.3. Image annotation

4.3.1. Tag completion

In the experiments of tag completion, we first evaluate the performance of our proposed approach on the training datasets that are partially annotated. In particular, similar to Lin et al. (2013), we randomly remove 40% of tags annotated in all training and testing images. Moreover, we measure our proposed approach and competitors in terms of AP@N and AR@N, where N is the number of top completed tags. For IAPR-TC12 and NUS-WIDE, N is varied with respect to 1,3,5. For LabelMe, since the mean number of tags per image in this dataset is higher than the others, the mean number of missing tags is also higher. Then, we vary N with respect to 2,4,6. Fig. 7 (a)–(f) shows the results for three datasets. We can observe that LSR and our proposed approach outperform the remaining approaches in all cases. Moreover, the performance of our proposed approach is slightly better than LSR except for AP@N on NUS-WIDE. We also observe that the tag completion algorithms (e.g. LSR, TMC) obtain significantly better performance than the automatic image annotation algorithms (e.g. TagProp, TagNV), since these algorithms are designed on the training data with clean annotation. For topic model-based algorithms, FeaCorrLDA outperforms CorrLDA by a large margin in most cases due to its improvement of topic extraction as we shown in Section 4.2.

We further evaluate the performance of tag completion on the training images that are fully annotated. This allows us to examine how well our proposed approach performs on a more idealistic case of tag completion. In this experiment, we remove 40% of tags annotated in only testing images. Hence, the mean number of missing tags is 3 for IAPR-TC12 and NUS-WIDE datasets, and 4 for LabelMe dataset. Table 4 presents the experimental results on three datasets. We observe that FeaCorrlDA still obtains superior performance than the other approaches. This shows the effectiveness of our proposed approach when dealing with various cases of training data.

Table 4
Tag completion results on the training images that are fully annotated.

	LabelMe		IAPR-TC12		NUS-WIDE	
	AP@4	AR@4	AP@3	AR@3	AP@3	AR@3
TagProp	0.416	0.429	0.307	0.418	0.163	0.171
TagNV	0.394	0.413	0.282	0.377	0.146	0.168
TMC	0.405	0.391	0.312	0.406	0.187	0.226
LSR	0.44	0.477	0.317	0.449	0.221	0.293
CorrLDA	0.389	0.408	0.296	0.378	0.17	0.199
FeaCorrLDA	0.488	0.532	0.335	0.466	0.226	0.311

Table 5
Image annotation results on the training images that are fully annotated by human annotation.

	LabelMe		IAPR-TC12		NUS-WIDE	
	AP@8	AR@8	AP@6	AR@6	AP@6	AR@6
TagProp	0.303	0.427	0.483	0.317	0.295	0.141
TagNV	0.291	0.394	0.454	0.292	0.246	0.113
TMC	0.332	0.424	0.473	0.305	0.285	0.146
CorrLDA	0.314	0.418	0.462	0.280	0.251	0.115
FeaCorrLDA	0.384	0.493	0.508	0.347	0.331	0.183

4.3.2. Automatic image annotation

We then examine the performance of automatic image annotation where no tags are annotated in testing images. Here we conduct experiments on the training images that are completely annotated. Then, there are two cases of tag completion: 1) all training images are fully labeled by human annotation, which correspond to the ground-truth of benchmark datasets; 2) the training images are fully annotated by running the tag completion algorithm. In this case, we utilize FeaCorrLDA to complete the tags of the training images, because it yields the best performance as shown in previous experiments of tag completion. Moreover, we also want to test the effectiveness of our proposed approach when comparing with perfect case of manual annotation.

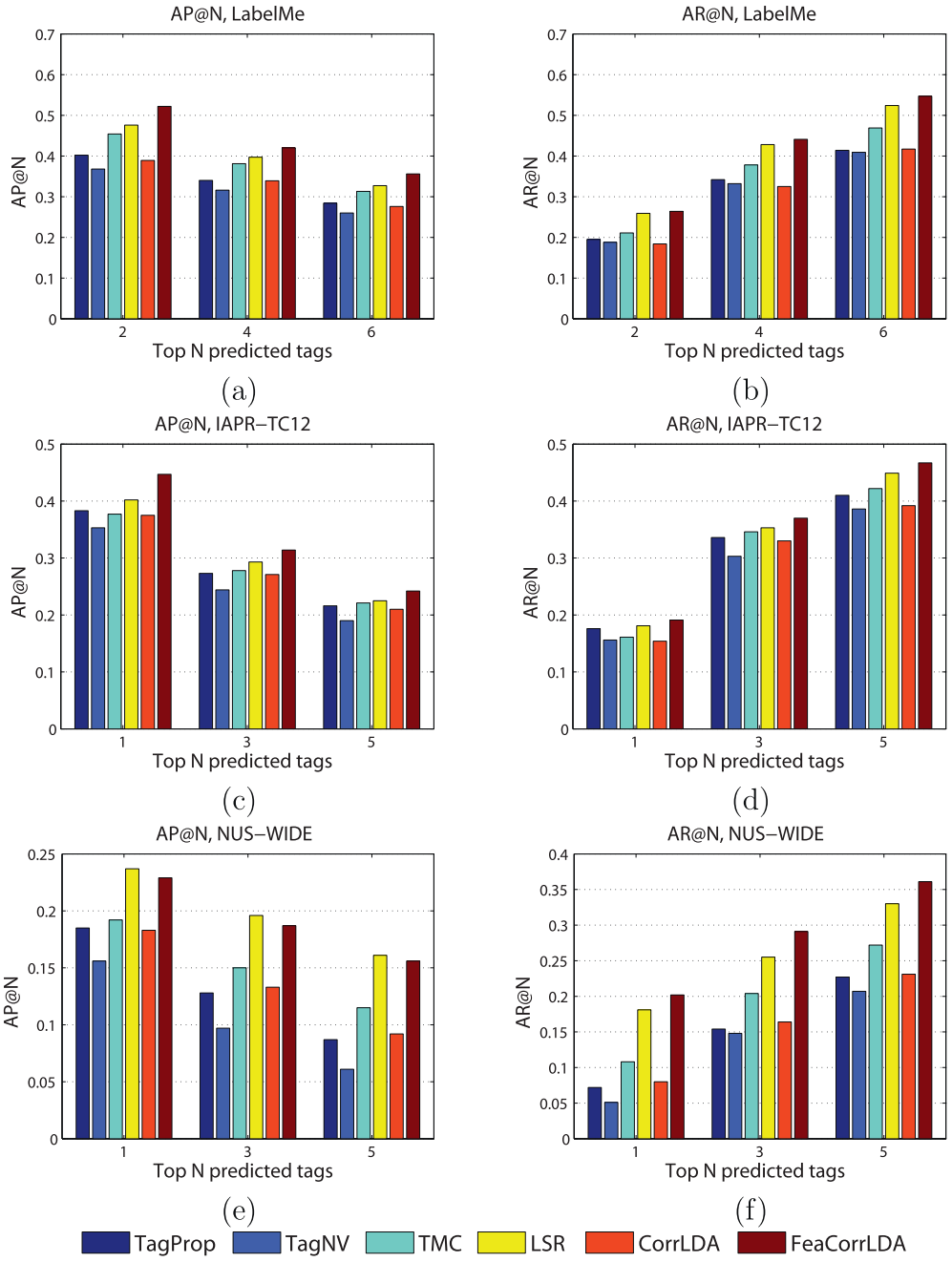

Fig. 7. Tag completion results in terms of AP@N and AR@N on the training images that are partially annotated.

Table 5 shows the results of image annotation on the testing images with the training datasets manually labeled. Note that we cannot perform automatic image annotation by using LSR, because this approach requires each testing image should be at least labeled by one tag. This is one of the major limitations of LSR where it is only suitable for the task of tag completion. In addition, two learning categories are employed in this experiments including transductive learning and inductive learning. In particular, TMC uses transductive learning to perform tag completion by including both training and testing images into the tag matrix. Hence,this approach has no distinction between training and testing sets. Unlike TMC, FeaCorrLDA and the remaining approaches utilize inductive learning where we process training and testing sets independently. While the transductive learning may obtain high performance, the inductive one has the better scalability to deal with real-world applications. In addition, although we can adopt transductive learning for FeaCorrlDA, the scalability is also the major focus of this study. For this reason, we prefer to use inductive learning in this experiment. As shown in Table 5, we can see that our proposed approach outperforms the others, even the transductive approach like TMC, in terms of both AP and AR on three datasets. This further shows the effectiveness and robustness of our algorithm.

Table 6 presents the experimental results on the training sets that are fully annotated by FeaCorrLDA. This experiment is only suitable for inductive methods where we treat training and testing sets independently. Therefore, we cannot obtain the results by using TMC, which only supports the transductive learning. Hence, we exclude TMC in this case. From Table 6, we also observe that our proposed approach yields superior performance in all datasets. In addition, the performance difference between the tag completion

Table 6
Image annotation results on the training images that are fully annotated by FeaCorrLDA.

	LabelMe		IAPR-TC12		NUS-WIDE	
	AP@8	AR@8	AP@6	AR@6	AP@6	AR@6
TagProp	0.281	0.397	0.425	0.263	0.275	0.13
TagNV	0.286	0.354	0.386	0.256	0.196	0.103
CorrLDA	0.274	0.373	0.403	0.259	0.217	0.125
FeaCorrLDA	0.353	0.459	0.469	0.307	0.301	0.164

Table 7
mAP for image retrieval using single keyword queries.

	LabelMe	IAPR-TC12	NUS-WIDE
TagProp	0.724	0.592	0.495
TagNV	0.694	0.583	0.473
TMC	0.745	0.614	0.518
LSR	0.775	0.648	0.533
CorrLDA	0.726	0.581	0.502
FeaCorrLDA	0.802	0.660	0.555

by human annotation and by FeaCorrLDA is not significant, where it only varies from 2% to 6% for AP, and 2% to 5% for AR. As a result, this experiment demonstrates the effectiveness of our FeaCorrLDA.

4.4. Semantic image retrieval

In the experiments of image retrieval, we treat each dataset consisting of both training and testing images as the database images. To guarantee the realistic setup of social image retrieval, all database images are partially annotated as the first experiment of tag completion in Section 4.3. Moreover, for each type of the experiment, we create a set of queries to evaluate the retrieval performance of different methods. In the following experiments, we show the results for two types of queries including keyword queries and image queries.

4.4.1. Image retrieval using keyword queries

We examine the retrieval performance on both single and multiple keyword queries. For single keyword queries, we consider each text word from the vocabulary of each dataset as an query. As a result, we have 490 queries for Labelme dataset, 291 queries for IAPR TC12 dataset, and 1000 queries for NUS-WIDE dataset. For multiple keyword queries, we create a set of queries composed of at least two text words. These combined words co-occur multiple times in the ground-truth database. Accordingly, we consider 126 queries for LabelMe dataset, 219 queries for IAPR-TC12 dataset, and 523 queries for NUS-WIDE dataset. Moreover, we perform the image annotation and tag completion methods on the partially annotated database images of each dataset to complete the tags of these images. Then, the relevance between a query and database images is verified if the image completely labeled contains all keywords of the query.

Table 7 shows the results of image retrieval using single keyword. We can see that obtained results are nearly consistent with the results of tag completion as shown in Section 4.3. In particular, our proposed method outperforms all the existing methods on three datasets. FeaCorrlDA is followed by LSR to yield the high performance. Otherwise, similar to the previous experiments, the tag completion algorithms perform significantly better than the image annotation algorithms.

The results of image retrieval using multiple keywords are presented in Table 8. Similar to the experiment of single keyword queries, our FeaCorrLDA also outperforms the other methods significantly for all three datasets. Hence, we can conclude that the retrieval performance of keyword-based image retrieval totally reflects the effectiveness of tag completion or image annotation algorithms.

Table 8
mAP for image retrieval using multiple keyword queries.

	LabelMe	IAPR-TC12	NUS-WIDE
TagProp	0.619	0.510	0.388
TagNV	0.584	0.507	0.371
TMC	0.643	0.535	0.391
LSR	0.661	0.543	0.430
CorrLDA	0.601	0.492	0.368
FeaCorrLDA	0.692	0.577	0.448

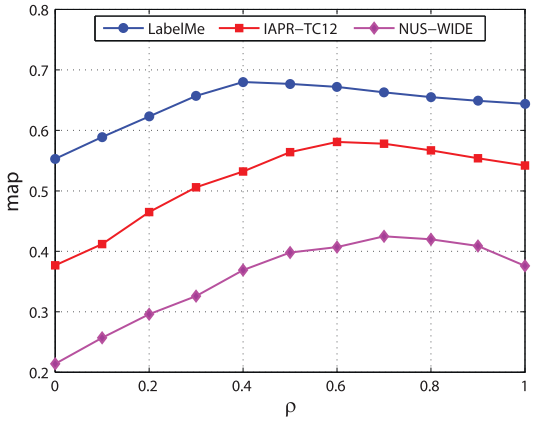

Fig. 8. Influence of weighting parameter ρ in term of mAP on three datasets.

4.4.2. Image retrieval using image queries

In this experiment, we further study another type of searching based on image queries. To create the queries, we randomly select a set of images from each dataset. Specifically, we collect 100 queries for Labelme dataset, 200 queries for IAPR TC12 dataset, and 500 queries for large-scale NUS-WIDE dataset. Then, for each query image, we manually check and select a number of tags that are most relevant to the content of the image. To make the ground-truth, we determine the relevance by checking whether the contents of database images that are fully labeled by human annotation reflect all the selected tags of query image. This rule is similar to the one of keyword-based retrieval, because it guarantees the reliability of semantic similarity between two images.

For the existing methods, the procedure to retrieve a list of relevant images is quite similar to the procedure of image annotation. In particular, given an image with no tags, we predict a number of tags reflecting the image content. Then, these tags are used to estimate the similarity between two images by using Eq. (14) with $\rho = 1$, which corresponds to the textual similarity. Unlike previous works, we consider the visual similarity between two images to guarantee the reliable search, where we use Eq. (14) with $0 < \rho < 1$. We further examine the influence of the weighting parameter ρ for the fusion of visual representation and textual tags. In this experiment, the retrieval performance (mAP) is measured for three datasets by varying ρ from 0 to 1. As shown in Fig. 8, the optimal values of setting ρ are 0.4, 0.6, and 0.7 for LabelMe, IAPR-TC12, and NUS-WIDE datasets, respectively. We can see that the optimally combined similarity outperforms both visual similarity ($\rho = 0$) and textual similarity ($\rho = 1$). This, therefore, proves the benefit of combining visual and textual similarities. Moreover, it can be observed that textual similarity tends to be weighted more on IAPR-TC12 and NUS-WIDE datasets. The reason is that these large-scale datasets have the high variety of data content, where the retrieved images relevant to a given query would be more semantically similar than visually similar.

Table 9
mAP for image retrieval using image queries.

	LabelMe	IAPR-TC12	NUS-WIDE
TagProp	0.584	0.492	0.334
TagNV	0.536	0.464	0.292
CorrLDA	0.558	0.486	0.326
FeaCorrLDA	0.644	0.542	0.376
ρ-FeaCorrLDA (BoV)	0.663	0.550	0.394
ρ-FeaCorrLDA (Fisher)	0.667	0.574	0.421
ρ-FeaCorrLDA (ColorFisher)	0.680	0.581	0.425

Table 10
mAP for image retrieval using image combined keyword queries.

	LabelMe	IAPR-TC12	NUS-WIDE
TagProp	0.644	0.559	0.419
TagNV	0.628	0.519	0.391
LSR	0.682	0.615	0.457
CorrLDA	0.629	0.543	0.404
FeaCorrLDA	0.726	0.660	0.465
ρ-FeaCorrLDA (BoV)	0.731	0.668	0.468
ρ-FeaCorrLDA (Fisher)	0.742	0.683	0.476
ρ-FeaCorrLDA (ColorFisher)	0.751	0.685	0.481

The retrieval results are shown in Table 9 on three datasets. Note that this experiment is only suitable for inductive learning where the query image is treated independently. Then, we cannot use TMC in this experiment. Otherwise, since image queries have no tags, we also cannot use LSR here. In Table 9, we denote our proposed methods as FeaCorrLDA and ρ-FeaCorrLDA, where FeaCorrLDA uses only textual similarity while ρ-FeaCorrLDA combines both textual and visual similarities with ρ selected as an optimal setting. Specifically, to compute the visual similarity with ρ-FeaCorrLDA, we perform three variants by using different types of image representation including BoV (Philbin et al., 2007), Fisher Vector (Sánchez et al., 2013), and our improved version of Fisher Vector called ColorFisher as presented in Section 3.5. We can observe that our proposed methods outperform the image annotation methods by a large margin. Moreover, ρ-FeaCorrLDA yields significantly better performance than FeaCorrLDA, which demonstrate the robustness of combining visual similarity. Otherwise, among three variants of ρ-FeaCorrLDA, BoV is worse than the others, while ColorFisher perform better than Fisher Vector. This, therefore, shows the effectiveness and our improvement when integrating color information into the Fisher Vector scheme.

We further evaluate the retrieval performance when using image queries combined with keywords. Here, we use the same set of queries for the experiment of queries using the only image. However, for each query image, we include a part of selected tags corresponding to that image as the searching keywords. In all three datasets, the number of keywords is varied from 2 to 5. We then use all the methods of former experiments for comparison. Since the image queries are partially labeled, we also include LSR in this experiment. Table 10 presents the results on three datasets. Similar to previous experiments, our proposed methods significantly outperform the others. Moreover, ColorFisher also performs better among the variants of ρ-FeaCorrLDA. On the other hand, tag completion algorithm like LSR yields significantly better performance than image annotation algorithms (e.g. TagProp, TagNV) due to its superior effectiveness of tag completion. However, LSR still performs worse than our FeaCorrLDA, which uses only textual similarity. Compared to the results shown in Table 9, in all cases, the performance of image combined keywords is better than the one of using the only image about 5–9%. Therefore, in practice, we can utilize keywords as a way to refine search results of image queries, and so increase the chance to retrieve more relevant images.

Insights into proposed approach: Through experimental results, we have ascertained some facts about the effectiveness of proposed approach. First, compared to the well-known CorrLDA model, the learning ability of our FeaCorrLDA model is significantly improved by integrating the visual characteristics (i.e., spatial location and color) of local features. Such an improvement positively affects the topic extraction and helps users to better interpret the latent topics which are highly consistent with semantic concepts as seen in Table 3. Moreover, benefiting from extracted topics, our model captures well the semantic association between visual features and tags, therefore, it is robust to the noise and incompleteness of tags in social images. The experimental results of tag completion and image annotation have further proved the robustness of our model by outperforming many state-of-the-art approaches on all datasets. We also note that results of tag prediction strongly affect the retrieval performance. This reveals the fact that better tag prediction increases the chances of retrieving more semantically relevant images. On the other hand, considering visual information in similarity estimation can be further effective for retrieval task using image queries. This is because the visual correspondence between the query and database images can be greatly enforced when the textual information is too personalized or general. We reflected this fact by showing our superior performance in Tables 9 and 10.

5. Conclusion

In this paper, we developed a new framework for intelligent image retrieval system via a learning approach that leverages user-provided tags from social media. These tags can help to bridge the semantic gap and facilitate the reliable search, but many of them are noisy and incomplete in describing the image contents. Consequently, we addressed the problem of tag completion and image annotation by proposing a probabilistic topic model. In particular, our model formulated the relationship among visual features as well as explicitly exploited the correlation of tags and image contents to extract the latent semantic topics, and enable the accurate tag prediction. Moreover, due to the generative nature, our model does not require the well-annotated images during the training process, hence efficiently dealing with large-scale data. We further proposed a generalized scoring scheme that handles well various query types by combining textual and visual similarities. Specifically, textual similarity was estimated based on the extracted topics while visual similarity was computed by the visual representations.

To validate our approach, we conducted experiments on three benchmark datasets: Labelme, IAPR TC12, and NUS-WIDE. We first evaluated our FeaCorrLDA model with the experiments of topic learning, which yielded much better results than conventional CorrLDA model. We then performed the experiments of tag completion and image annotation. In most cases, our method outperformed the existing methods for these tasks on three datasets. Finally, the experiments of image retrieval using different query types showed the superior performance of our model along with proposed scoring scheme when compared to the other methods. In summary, the obtained results demonstrated the great potential of our contributions for real-world applications.

For future works, the proposed approach can be extended in possible directions as follows: (1) combining topic models with intelligent image annotation techniques based on deep learning to further improve the robustness of retrieval system, (2) automatically determining the optimal number of topics using the nonparametric model (Teh, Jordan, Beal, & Blei, 2006) rather than current parametric models like CorrLDA and FeaCorrLDA where this number is empirically selected, (3) incorporating the temporal information into topic model to handle streaming and video data, (4) ex-

ploring the use of different feature types available in the literature to find which one is the best for annotation and retrieval tasks, (5) adopting efficient optimization procedure to learn the weighting parameter ρ of our scoring scheme and better manage the balance between textual and visual similarities.

Acknowledgment

This research was supported by the MSIP, Korea, under the G-ITRC support program (IITP-2016-R6812-16-0001) supervised by the IITP.

References

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning, 50*(1–2), 5–43.

Binder, A., Samek, W., Müller, K.-R., & Kawanabe, M. (2013). Enhanced representation and multi-task learning for image annotation. *Computer Vision and Image Understanding, 117*(5), 466–478.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 127–134). ACM.

Bundschus, M., Yu, S., Tresp, V., Rettinger, A., Dejori, M., & Kriegel, H.-P. (2009). Hierarchical bayesian models for collaborative tagging systems. In *Proceedings of the 9th IEEE international conference on data mining (ICDM)* (pp. 728–733). IEEE.

Cao, X., Zhang, H., Guo, X., Liu, S., & Meng, D. (2015). Sled: Semantic label embedding dictionary representation for multilabel image annotation. *IEEE Transactions on Image Processing, 24*(9), 2746–2759.

Carneiro, G., Chan, A. B., Moreno, P. J., & Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 394–410.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: A real–world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval* (p. 48). ACM.

Desai, C., Ramanan, D., & Fowlkes, C. C. (2011). Discriminative models for multi-class object layout. *International Journal of Computer Vision, 95*(1), 1–12.

Ding, G., Guo, Y., & Zhou, J. (2014). Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE international conference on computer vision and pattern recognition (cvpr)* (pp. 2075–2082).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Guillaumin, M., Mensink, T., Verbeek, J., & Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the 12th IEEE international conference on computer vision* (pp. 309–316). IEEE.

Hariharan, B., Zelnik-Manor, L., Varma, M., & Vishwanathan, S. (2010). Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 423–430).

Jiang, W., Chang, S.-F., & Loui, A. C. (2006). Active context-based concept fusionwith partial user labels. In *Proceedings of the IEEE international conference on image processing* (pp. 2917–2920). IEEE.

Krestel, R., & Fankhauser, P. (2009). Tag recommendation using probabilistic topic models. *ECML PKDD Discovery Challenge, 2009*, 131.

Lee, S., De Neve, W., & Ro, Y. M. (2010). Tag refinement in an image folksonomy using visual similarity and tag co-occurrence statistics. *Signal Processing: Image Communication, 25*(10), 761–773.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications, 2*(1), 1–19.

Li, X., Snoek, C. G., & Worring, M. (2009). Learning social tag relevance by neighbor voting. *IEEE Trans. on Multimedia, 11*(7), 1310–1322.

Lin, Z., Ding, G., Hu, M., Wang, J., & Ye, X. (2013). Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)* (pp. 1618–1625). IEEE.

Liu, D., Hua, X.-S., Wang, M., & Zhang, H.-J. (2010). Image retagging. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 491–500). ACM.

Liu, Y., Wu, F., Zhang, Y., Shao, J., & Zhuang, Y. (2011). Tag clustering and refinement on semantic unity graph. In *Proceedings of the 11th IEEE international conference on data mining (ICDM), 2011* (pp. 417–426). IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Monay, F., & Gatica-Perez, D. (2004). Plsa-based image auto-annotation: constraining the latent space. In *Proceedings of the 12th annual ACM international conference on multimedia* (pp. 348–351). ACM.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)* (pp. 1–8). IEEE.

Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., & Zhang, H.-J. (2007). Correlative multi-label video annotation. In *Proceedings of the 15th international conference on multimedia* (pp. 17–26). ACM.

Rasiwasia, N., & Vasconcelos, N. (2013). Latent dirichlet allocation models for image classification. *IEEE Tranactions on Pattern Analysis and Machine Intelligence, 35*(11), 2665–2679.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1–3), 157–173.

Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision, 105*(3), 222–245.

Song, J., Yang, Y., Yang, Y., Huang, Z., & Shen, H. T. (2013). Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 785–796). ACM.

Teh, Y. W., Jordan, M., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581.

Tu, N. A., Cho, J., & Lee, Y.-K. (2016). Semantic image retrieval using correspondence topic model with background distribution. In *Proceedings of the 2016 international conference on big data and smart computing (BIGCOMP)* (pp. 191–198). IEEE.

Van De Sande, K. E., Gevers, T., & Snoek, C. G. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1582–1596.

Van De Weijer, J., & Schmid, C. (2006). Coloring local feature extraction. In *Proceedings of European conference on computer vision (ECCV) 2006* (pp. 334–348). Springer.

Wang, C., Blei, D., & Li, F.-F. (2009). Simultaneous image classification and annotation. In *Proceedings of the IEEE international conference on computer vision and pattern recognition (cvpr)* (pp. 1903–1910). IEEE.

Wang, C., Jing, F., Zhang, L., & Zhang, H.-J. (2006). Image annotation refinement using random walk with restarts. In *Proceedings of the 14th annual ACM international conference on multimedia* (pp. 647–650). ACM.

Wang, J., Zhou, J., Xu, H., Mei, T., Hua, X.-S., & Li, S. (2014). Image tag refinement by regularized latent dirichlet allocation. *Computer Vision and Image Understanding, 124*, 61–70.

Weinberger, K. Q., Slaney, M., & Van Zwol, R. (2008). Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on multimedia* (pp. 111–120). ACM.

Wu, L., Jin, R., & Jain, A. K. (2013). Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(3), 716–727.

Wu, P., Hoi, S. C.-H., Zhao, P., & He, Y. (2011). 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* (pp. 197–206). ACM.

Xia, Z., Feng, X., Peng, J., Wu, J., & Fan, J. (2015). A regularized optimization framework for tag completion and image retrieval. *Neurocomputing, 147*, 500–508.

Yang, Y., Xu, D., Nie, F., Luo, J., & Zhuang, Y. (2009). Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM international conference on multimedia* (pp. 175–184). ACM.

Yang, Y., Zhuang, Y.-T., Wu, F., & Pan, Y.-H. (2008). Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia, 10*(3), 437–446.

Zhou, N., Cheung, W. K., Qiu, G., & Xue, X. (2011). A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(7), 1281–1294.

Zhu, G., Yan, S., & Ma, Y. (2010). Image tag refinement towards low-rank, content–tag prior and error sparsity. In *Proceedingsof the ACM international conference on multimedia* (pp. 461–470). ACM.

Zhuang, J., & Hoi, S. C. (2011). A two-view learning approach for image tag ranking. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 625–634). ACM.

Zhuang, Y.-T., Yang, Y., & Wu, F. (2008). Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia, 10*(2), 221–229.