A Bayesian Approach to Multimodal Visual Dictionary Learning

Go Irie[†], Dong Liu[‡], Zhenguo Li[‡], Shih-Fu Chang[‡]
[†] NTT Corporation, Kanagawa, Japan
[‡] Columbia University, New York, USA
[†]irie.go@lab.ntt.co.jp [‡]{dongliu, zgli, sfchang}@ee.columbia.edu

Abstract

Despite significant progress, most existing visual dictionary learning methods rely on image descriptors alone or together with class labels. However, Web images are often associated with text data which may carry substantial information regarding image semantics, and may be exploited for visual dictionary learning. This paper explores this idea by leveraging relational information between image descriptors and textual words via co-clustering, in addition to information of image descriptors. Existing co-clustering methods are not optimal for this problem because they ignore the structure of image descriptors in the continuous space, which is crucial for capturing visual characteristics of images. We propose a novel Bayesian co-clustering model to jointly estimate the underlying distributions of the continuous image descriptors as well as the relationship between such distributions and the textual words through a unified Bayesian inference. Extensive experiments on image categorization and retrieval have validated the substantial value of the proposed joint modeling in improving visual dictionary learning, where our model shows superior performance over several recent methods.

1. Introduction

Image representation is a starting point for visual understanding and retrieval. The histogram of (quantized) local image descriptors like bag-of-visual-words (BoVW) is the most popular image representation in computer vision. The process is to first train a visual dictionary based on the extracted descriptors from an image collection and then to encode the descriptors of each image into a histogram based on the learned dictionary. The crucial issue is how to build a visual dictionary.

There has been considerable interest in visual dictionary learning which can be classified into two paradigms. The first paradigm is unsupervised learning such as K-means for BoVW and LLC [25] for sparse representation. However, these methods do not have any mechanism to retain discriminative information (e.g., object or scene categories) in visual dictionaries. The second one is the supervised learning, which incorporates class labels into the visual dictionary [12, 30, 9, 10, 17]. However, it generally needs manual labors to annotate class labels, which are difficult, if not impossible, to obtain in many real-world problems.

Regardless of whether unsupervised or supervised, existing visual dictionary learning methods are based on only *single-modal information*, i.e., information of image descriptors alone. On the other hand, besides the plain visual content, a huge number of Web images are augmented with text descriptions. Photos in Flickr are frequently accompanied by tags and each picture in Wikipedia is associated with a document describing its semantic topics. Such text information is generally noisy but provides semantic cues about the image content, facilitating the design of multimodal system for image understanding. A recent work in [7] combines images and texts through multiple kernel learning, and demonstrates that text information can significantly improve the performance of image classification.

These facts motivate us to consider leveraging textual words for visual dictionary learning. Specifically, the problem can be stated as following: given a set of images and their associated textual words, how to learn a visual dictionary that incorporates both image and text information. This is challenging due to the following two key issues. First, there are a large number of local descriptors extracted from the images, whose corresponding relations to the textual words are totally unknown, making it difficult to explore the multimodal correlation. Second, the visual and textual spaces are completely different from each other: image descriptors are generally in a *continuous* space (e.g., SIFT is typically represented as a 128-dimensional real-valued vector) whereas textual words are in a *discrete* space.

Addressing these issues, we propose a novel approach for learning a visual dictionary from both image and text information. The whole framework is illustrated in Fig. 1. Our approach is inspired by the *co-clustering* framework. Given a relational (co-occurrence) matrix between image descriptors and textual words, we perform clustering along the image descriptors and textual words simultaneously and obtain disjoint clusters for each modality, in which the image descriptor clusters can be used as a visual dictionary while the textual word clusters reveal the semantic topics of the entire image collection. The clusters of image descriptors are determined based on their relations with respect to the textual word clusters, which well captures the multimodal correlation. Note that the textual word clusters are important for discovering the significant multimodal correlation, due to the fact that the individual word is noisy and may not convey beneficial information while the clusters of multiple words can reflect the semantic topic of the consti-

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.49

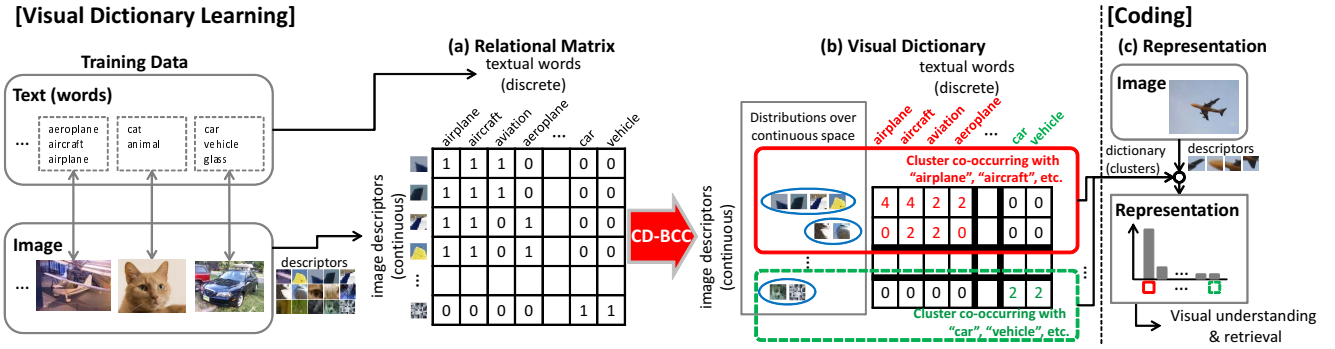


Figure 1. Illustration of the proposed framework of multimodal visual dictionary learning. (a) Given a set of training images and associated textual words, we first form a relational matrix that represents the relationship between image descriptors (rows) and textual words (columns), where each element is 1 if the corresponding pair is extracted from an identical image and 0 otherwise. (b) Our continuous-discrete Bayesian co-clustering (CD-BCC) jointly estimates distributions of the continuous image descriptors as well as the relationship between the image descriptor distributions and textual words. During a unified Bayesian inference process, a subset of image descriptors from an identical distribution (represented by an ellipse) is aggregated into a single row, and the rows (distributions) are further grouped into fewer number of clusters based on their co-occurrence frequencies with textual words. Each resulting cluster of image descriptors is thus expected to (i) have consistently co-occurring textual words and (ii) be visually different from the other clusters. (c) These clusters form the final visual dictionary and are used to encode an image into a single image representation vector (histogram).

tutive words more sufficiently.

Specifically, we investigate how to perform co-clustering along a *continuous* image descriptor space and a *discrete* textual word space simultaneously, and propose continuous-discrete Bayesian co-clustering (CD-BCC). Previous co-clustering methods [3, 4, 5, 22] may not be optimal for visual dictionary learning because they perform clustering relying on only the relational information of image descriptors and textual words, which do not have a potential to incorporate any visual properties into the visual dictionary. A straightforward approach may be first to quantize image descriptors using K-means and then to co-cluster quantized descriptors (visual words) and textual words. However, such two step approach is inevitably subject to information loss and may degrade the performance. Unlike these, our CD-BCC simultaneously estimates the underlying distributions of image descriptors over the continuous space and the relationship between the distributions and textual words via a unified Bayesian inference framework. Consequently, each image descriptor cluster used to construct each dimension of the final image representation vector is ideally consistent to a set of textual words with consistent semantic topic as well as visually different from the other clusters.

Extensive experiments on five different datasets will demonstrate that the proposed multimodal visual dictionary learning approach can achieve significant performance gains when evaluated over various tasks including image classification and content-based image retrieval (CBIR). We will also show clear evidences confirming the capabilities of the method for capturing the multimodel information in the learned visual dictionary.

2. Related Work

We review some recent studies on visual dictionary learning, co-clustering, and multimodal topic modeling.

Visual Dictionary Learning: Standard unsupervised methods like K-means, K-SVD [1], and LLC [25] train visual dictionaries based on only information of images. There are many supervised methods aiming at retaining discriminative information in visual dictionaries [12, 28, 31, 30, 9, 10, 17, 16]. [12] trains a dictionary so as to maximize the mutual information. [28, 31] learns multiple class-specific dictionaries. [30, 9, 10, 17, 16] learn a single visual dictionary by jointly optimizing visual dictionaries and discriminative functions. We aim to leverage *weak* textual words associated with images, instead of assuming *strong* class labels. Prior supervised methods assume that class labels are mutually exclusive, which may not be reasonable in our problem because there is often strong correlation among textual words. Contrary, our approach is based on co-clustering and takes into account textual word clusters which may effectively guide visual feature clustering. Also ours is different from multimodal image classifier learning methods [7, 20] from the same view point.

Co-clustering: Co-clustering is an emerging paradigm that is often used to cluster two types of variables simultaneously under given a relational matrix. Many co-clustering methods have been proposed. Spectral co-clustering (SCC) [3] solves the problem as spectral partitioning over a bipartite graph. Information theoretic co-clustering (ITCC) [4] determines clusters so as to minimize the loss of mutual information between a given relational matrix and its co-clustering results. Non-negative matrix tri-factorization (NMTF) [5] decomposes a matrix into three non-negative matrices, where two of them correspond to row and column clusters respectively. The most relevant approach to ours is Bayesian co-clustering (BCC) [22] which is a generative model of a relational matrix and estimates clusters of rows and columns in a Bayesian inference framework. Several extensions have also been developed so far [26, 18, 11].

Unlike these previous methods, our CD-BCC is for a pair of continuous and discrete variables and jointly estimates the distributions of image descriptors over the continuous space and co-clusters of the distributions and textual words. To the best of our knowledge, this is the first work that presents a co-clustering method for continuous-discrete variable pairs.

Multimodal Topic Models: Topic modeling (e.g., latent semantic analysis (LSA) and latent Dirichlet allocation (LDA)) has become a popular paradigm to jointly model multiple information sources. For instance, [29, 32] proposed LSA-based methods to model visual and textual words with an underlying latent topic space. [21] proposed a cross-modal retrieval approach based on LDA with canonical correlation analysis. More related approaches to ours will be Bayesian multimodal topic models [2, 15, 24, 19, 8, 14]. In particular, Li et al [14] proposed a Bayesian multimodal topic model for visual dictionary learning. Our CD-BCC is also a Bayesian model for visual dictionary learning – but ours is a co-clustering model, not a topic model. Specifically, topic models assume some mixture distributions over visual words as well as textual words. Our CD-BCC also assumes a mixture over image descriptors, but does not assume any distributions over textual words. Instead, we assume a mixture over a relational matrix that encourages the model to identify the significant multimodal correlation from sparse and noisy relational data.

3. Multimodal Visual Dictionary Learning

We assume the typical image descriptor extraction process: a set of key points are detected from each training image first, and then an image descriptor (e.g., SIFT) is extracted from each key point. Let us introduce some notations on observation variables first. Suppose we have N image descriptors $\mathbf{X} = \{x_1, \dots, x_N\}$, where $x_i \in \mathbb{R}^d$ is a descriptor extracted from an training image, and (at least a subset of) the training images are associated with words, where each word is one entity from a vocabulary indexed by $\{1, \dots, W\}$. Assuming that a pair of an image descriptor and a word from the same image are "related", the relationship between N image descriptors and W unique words is naturally represented as an $N \times W$ initial relational matrix $\mathbf{R} \equiv \{r_{i,j}\}$, where $r_{i,j} = 1$ if i-th image descriptor and j-th word are related and 0 otherwise. Note that each row corresponds to each image descriptor, thus the i-th row is associated with a descriptor $x_i \in \mathbb{R}^d$. Then our problem is: given the initial relational matrix $\mathbf{R}$ and the corresponding set of image descriptors $\mathbf{X}$, the goal is to find K clusters of image descriptors $\mathbf{X}$ used to form a visual dictionary with the assistance of the relational matrix $\mathbf{R}$.

3.1. Visual Dictionary Learning with Continuous-Discrete Bayesian Co-clustering

We propose CD-BCC to solve the problem. The motivation of the model is to find out unknown K image descriptor clusters depending on both of the distributions of image descriptors $\mathbf{X}$ over a continuous space and their relationship with respect to the words presented in the relational matrix $\mathbf{R}$. To achieve this, CD-BCC is designed as a Bayesian generative model, i.e., a joint distribution of image descriptors $\mathbf{X}$, the relational matrix $\mathbf{R}$, and image descriptor clusters. Based on the joint distribution, the image descriptor clusters are estimated based on its Bayesian posterior distributions under given $\mathbf{X}$ and $\mathbf{R}$. We first introduce the generative process of our CD-BCC, and then provide the inference scheme to estimate the posterior distributions.

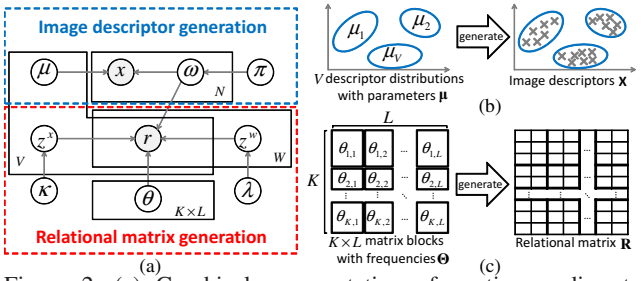


Figure 2. (a) Graphical representation of continuous-discrete Bayesian co-clustering (CD-BCC), where the hyperparameters are omitted for simplicity. Illustrations of generative processes: (b) image descriptor generation and (c) relational matrix generation.

3.1.1 Generative Process

Fig. 2 shows a graphical representation of our CD-BCC. In CD-BCC, $\mathbf{X}$ and $\mathbf{R}$ are generated via seamlessly integrated two sub-processes. In the *image descriptor generation* process, all image descriptors $\mathbf{X}$ are generated from a mixture of V *descriptor distributions* over the continuous image descriptor space. In the *relational matrix generation* process, we consider a (new) relational matrix $\mathbf{R}$ that represents the co-occurrence between each pair of V descriptor distributions and W words, and assume that the relational matrix $\mathbf{R}$ is generated from $K \times L$ co-occurrence blocks, each of which is a co-cluster of descriptor distributions and words and measures the mutual dependency among the constitutive image descriptors and words.

Mathematically, the each sub-process is described as follows. Without loss of generality, we assume that $\mathbf{X}$ are sifted so as to be zero-mean hereafter.

Image descriptor generation: N image descriptors $\mathbf{X}$ are generated from a mixture of V descriptor distributions $\text{Norm}(\mu_v, \Sigma_x)$ $(v = 1, \dots, V)$ as the following process:

1. For each descriptor distribution v, draw mean $\mu_v; \Sigma_0 \sim \text{Norm}(\mathbf{0}, \Sigma_0)$.
2. Draw mixture proportion of descriptor distributions $\pi; \gamma, V \sim \text{Dir}(\gamma/V)$.
3. For each image descriptor x_i
   (a) Draw descriptor distribution assignment $\omega_i; \pi \sim \text{Mult}(\pi)$
   (b) Draw image descriptor $x_i|\boldsymbol{\mu}, \omega_i; \Sigma_x \sim \text{Norm}(\mu_{\omega_i}; \Sigma_x)$

In Step 1, mean vectors of V descriptor distributions, $\boldsymbol{\mu}$, are generated. Step 2 generates the mixture proportion π that controls how much frequently each descriptor distribution is used. In Step 3, for each image descriptor x_i, (a) one descriptor distribution $\omega_i \in \{1, \dots, V\}$ is chosen with the

probability Mult(π), and then (b) x_i is generated from ω_i-th descriptor distribution Norm(μ_{ω_i}, Σ_x).

Relational matrix generation: The relational matrix $\mathbf{R}$ is generated from Poisson($\theta_{k,l}$) as the following process:

4. For each block (k, l) (i.e., each pair of image descriptor cluster and word cluster), draw co-occurrence frequency $\theta_{k,l}; \beta, \phi \sim$ Gamma(β, ϕ).
5. Draw image descriptor cluster proportion $\kappa; \zeta, K \sim$ Dir(ζ/K) and word cluster proportion $\lambda; \eta, L \sim$ Dir(η/L) respectively.
6. For each descriptor distribution v and each word j, draw image descriptor cluster assignment $z_v^x|\kappa \sim$ Mult(κ) and word cluster assignment $z_j^w|\lambda \sim$ Mult(λ) respectively.
7. For each pair of descriptor distribution and word (v, j), draw element of relational matrix $r_{v,j}|\mathbf{\Theta}, z_v^x, z_j^w \sim$ Poisson$(\theta_{z_v^x, z_j^w})$.

Step 4 generates $K \times L$ block co-occurrence frequencies $\mathbf{\Theta}$. The image descriptor cluster of v-th descriptor distribution, $z_v^x \in \{1, \ldots, K\}$, (and the word cluster of j-th word, $z_j^w \in \{1, \ldots, L\}$) is generated in Step 6. The (v, j)-th element of the relational matrix $\mathbf{R}$, $r_{v,j}$, is generated from Poisson$(\theta_{z_v^x, z_j^w})$ in Step 7. Note that $r_{v,j}$ denotes the number of times image descriptors in v-th descriptor distributions co-occurs with j-th word[1].

3.1.2 Visual Dictionary Inference

Observing the image descriptors $\mathbf{X}$ and the relational matrix $\mathbf{R}$, we compute the posterior distributions to infer the image descriptor clusters $\boldsymbol{z}^x$ and the mean vectors of the descriptor distributions $\boldsymbol{\mu}$ used as the visual dictionary. The above generative process determines the joint distribution $p(\mathbf{X}, \mathbf{R}, \boldsymbol{\omega}, \boldsymbol{z}^x, \boldsymbol{z}^w, \pi, \kappa, \lambda, \boldsymbol{\mu}, \mathbf{\Theta})$. Then what we want to know is the posterior $p(\boldsymbol{\omega}, \boldsymbol{z}^x, \boldsymbol{z}^w, \pi, \kappa, \lambda, \boldsymbol{\mu}, \mathbf{\Theta}|\mathbf{X}, \mathbf{R})$. However, this is computationally intractable. Fortunately, because CD-BCC is designed as a full conjugate model, samples from the posterior can be efficiently obtained by collapsed Gibbs sampler. Specifically, we integrate out π, κ, λ, $\boldsymbol{\mu}$, and $\mathbf{\Theta}$ from the joint distribution, and samples only $\boldsymbol{\omega}$, $\boldsymbol{z}^x$, and $\boldsymbol{z}^w$ using the marginal distribution $p(\mathbf{X}, \mathbf{R}, \boldsymbol{\omega}, \boldsymbol{z}^x, \boldsymbol{z}^w)$ at each iteration step (the actual sampling distributions are shown in Appendix). Regarding the computation time, one iteration typically takes around 10 minutes using 64-bit Matlab on a machine with 2.4 GHz Xeon CPU and 32GB RAM. After iterations, the mean vectors $\boldsymbol{\mu}$ of V descriptor distributions are estimated as $\mu_v = \sum_{i \text{ s.t. } \omega_i = v} x_i$.

Note that $\boldsymbol{\mu}$ determines the distributions of the image descriptors while $\boldsymbol{z}^x$ determines the clusters of these distributions based on their correlation to word clusters. Therefore, they can be used as the visual dictionary to generate image representation for the new images. In the next section, we will explain how to utilize them to generate textual information embedded image representation.

[1] Introducing an index matrix $\hat{\mathbf{\Omega}} = \{\hat{\omega}_{i,v}\}$ such that $\omega_{i,v} = 1$ if $\omega_i = v$ and 0 otherwise, this is efficiently computed as $\mathbf{R} = \hat{\mathbf{\Omega}}^\top \mathbf{R}_0$, where $\mathbf{R}_0$ is the initial relational matrix between N image descriptors and W words.

3.2. Encoding Methods

With the trained $\boldsymbol{\mu}$ and $\boldsymbol{z}^x$, a test image represented as a set of I descriptors $\{x_1, \ldots, x_q, \ldots, x_I\}$ is encoded into a K dimensional image-level representation c. In computer vision, dictionary learning and coding are seen as independent processes (e.g., K-means dictionaries are used for vector quantization, sparse coding [27], or LLC [25]). We here consider two possible coding approaches: maximum likelihood (ML) coding and sparse coding.

Maximum Likelihood (ML) Coding: One natural way is ML coding that uses $\boldsymbol{\mu}$ and $\boldsymbol{z}^x$ as "plug-in" estimators. Each image descriptor x_q is efficiently encoded into $z_q \in \{1, \ldots, K\}$ without any iterative inference:

$$\omega_q = \arg\max_v p(x_q|\mu_v; \Sigma_x) \quad (1)$$

$$z_q = z^x_{\omega_q} \quad (2)$$

The final image representation c is obtained by counting frequencies of $\{1, \ldots, K\}$ in $\{z_1, \ldots, z_q, \ldots, z_I\}$ (average-pooling).

Sparse Coding: Note that $\boldsymbol{\mu}$ and $\boldsymbol{z}^x$ can be regarded as a set of V basis vectors of $\mathbb{R}^d$ and a projection from $\{1, \ldots, V\}$ to $\{1, \ldots, K\}$ ($K \leq V$) respectively. Inspired by the fact, we apply sparse coding [27] with the dictionary $\boldsymbol{\mu}$ to generate an initial code $\boldsymbol{\omega}_q$ and then reduce its dimensionality:

$$\boldsymbol{\omega}_q = \arg\max_{\boldsymbol{\omega}} ||x_q - \boldsymbol{\omega}^\top \boldsymbol{\mu}||_2 + \lambda||\boldsymbol{\omega}||_1 \quad (3)$$

$$\boldsymbol{z}_q = \boldsymbol{\omega}_q^\top \mathbf{Z} \quad (4)$$

where $\mathbf{Z} \equiv \{z_{v,k}\}$ is a matrix such that $z_{v,k} = 1$ if $z_v^x = k$ and 0 otherwise. The final image representation c is obtained by applying max-pooling to $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q, \ldots, \boldsymbol{z}_I\}$.

To show the significance of $\boldsymbol{z}^x$, we provide examples of image representation vectors by ML coding in Fig. 3. The final image representations with both $\boldsymbol{z}^x$ and $\boldsymbol{\mu}$ are sparser than initial representations with only $\boldsymbol{\mu}$. Moreover, average l_2 distances between images of the same category (different categories) based on the final representations are smaller (larger) than those by initial representations. This can be a clear evidence that our CD-BCC successfully captures discriminative information from textual words via $\boldsymbol{z}^x$.

3.3. Non-parametric Extension

We note that we need to manually setup the following three parameters: the final dictionary size K, the number of descriptor distributions V and the number of word clusters L. As long as we are interested in only the final dictionary size K, the other two, V and L, are parameters. Adjusting such parameters often requires much effort because these are manually determined through a number of preliminary experiments. We therefore consider a non-parametric extension of CD-BCC to estimate these two parameters automatically based on the training data.

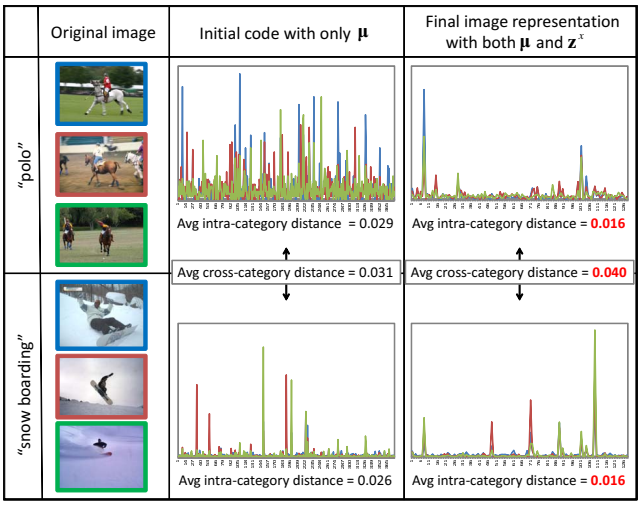


Figure 3. Examples of image representations encoded with the visual dictionary trained by our CD-BCC on UIUC-Sport dataset. Image representation vectors of three images from each of "polo" and "snow boarding" categories are shown. The third and fourth columns show initial and final image representations computed by Eq. (1) (377 dimensional) and Eq. (2) (128 dimensional) respectively. Different colors mean the results of different images. We also show average intra- and cross-category l_2 distances. This is best viewed in color.

Consider Step 2 and Step 3 (a) in the generative process. Let us assume a stochastic process $p(\omega_i|\omega_{1:i-1};\gamma)$ instead of Dir(γ/V) and Mult(π). Because these two are Dirichlet-multinomial conjugate pair, they can be immediately replaced by Chinese restaurant process (CRP) [23] by taking infinite limit $V \to \infty$ of $p(\omega_i|\omega_{1:i-1};\gamma)$:

$$p(\omega_i = t|\omega_{1:i-1};\gamma) \propto \begin{cases} \frac{m_t}{i-1+\gamma} & (m_t > 0) \\ \frac{\gamma}{i-1+\gamma} & (otherwise) \end{cases} \quad (5)$$

Similarly $p(z^w_j|z^w_{1:j-1};\eta)$ is also given as

$$p(z^w_j = b|z^w_{1:j-1};\eta) \propto \begin{cases} \frac{m^w_b}{j-1+\eta} & (m^w_b > 0) \\ \frac{\eta}{j-1+\eta} & (otherwise) \end{cases} \quad (6)$$

where m^w_b is the number of words assigned to b-th cluster. Based on these, collapsed Gibbs sampler for the non-parametric version is derived immediately (see Appendix). In all the experiments conducted in the next section, we consistently use this non-parametric version. Note that our CD-BCC always requires $V > K$. To satisfy it, we initialize $V = K$, and in case if $V < K$ during iterations then we sample $(K - V)$ new clusters at a time[2].

4. Experimental Results

We analyze the performance of our approach for the tasks of image categorization and CBIR using five different datasets. For all the experiments, we use only a single image descriptor for our CD-BCC, i.e., 128-dimensional grayscale SIFT extracted in a dense sampling manner, and fix hyper-parameters of our CD-BCC as $\gamma = 100$, $\eta = 5$, $\zeta = 5$, $\Sigma_0 = 0.1I$, $\Sigma_x = 0.1I$, $\beta = 1$, and $\phi = 1$ (other settings gave similar performance).

Table 1. Categorization accuracy on UIUC-Sport and LabelMe datasets. The results of [14, 24] are directly extracted from each paper. For the results of K-means, four co-clustering methods and Ours, average over 50 runs (different training/testing splits) are reported. $K = 256$.

	UIUC-Sport	LabelMe
K-means	65.2 ± 1.4	74.9 ± 0.9
Wang [24]	66	76
Li [14]	69.1	76.3
SCC [3]	74.7 ± 1.1	82.1 ± 0.8
ITCC [4]	72.9 ± 1.2	80.5 ± 0.8
NMTF [5]	62.9 ± 1.6	79.9 ± 1.1
BCC [22]	69.1 ± 1.6	82.5 ± 0.8
Ours	$\mathbf{75.1 \pm 1.4}$	$\mathbf{83.1 \pm 0.9}$

4.1. Image Categorization

We first analyze the image categorization performance on three datasets for event, scene, and object categorization tasks: UIUC-Sport, LabelMe, and Caltech101. The performance is measured by image categorization accuracy.

4.1.1 Evaluation on UIUC-Sport & LabelMe

We first evaluate our approach on UIUC-Sport[3] and LableMe[4]. These datasets are selected for direct comparison to a state-of-the-art Bayesian multimodal topic model for dictionary learning [14]. In addition, we compare our approach to single-modal (image only) K-means, another recent Bayesian multimodal topic model [24], and four co-clustering methods (SCC [3], ITCC [4], NMTF [5], and BCC [22]). These existing co-clustering methods require both variables to be discrete, so we first apply K-means based vector quantization to discretize image descriptors into $2K$ codewords, and then apply each co-clustering method to derive final dictionaries of size K. We follow the same setting as [14, 24]:

UIUC-Sport: This dataset contains 1579 images of 8 sports categories: "badminton" (200 images), "bocce" (137 images), "croquet" (236 images), "polo" (182 images), "rock climbing" (194 images), "rowing" (250 images), "sailing" (190 images), and "snow boarding" (190 images). We randomly split each class evenly and construct the training data and the testing data. Each image has 15 textual words on average. We select the 30 most frequent words ($W = 30$).

LabelMe: This dataset is a subset of LabelMe that has 8 image categories: "coast", "forest", "highway", "inside city", "mountain", "open country", "street" and "tall building". We randomly select 1600 images in total (200 images for each class), and randomly separate them to 800 training data and 800 testing data (100 images for each class in each

[2] Actually such a situation has never been observed through all the experiments conducted in this paper.

[3] http://vision.stanford.edu/lijiali/event_dataset/

[4] http://labelme.csail.mit.edu/

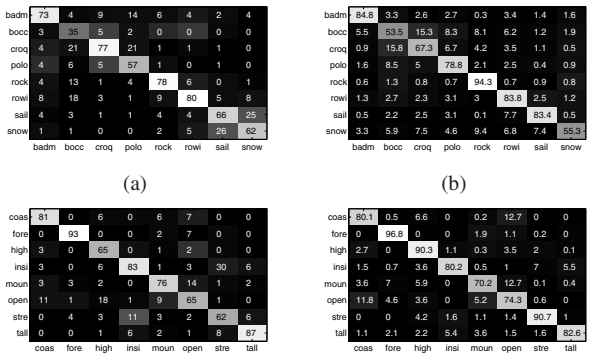


Figure 4. Confusion matrices: (a) [14] and (b) ours on UIUC-Sport, and (c) [14] and (d) ours on LabelMe.

Cluster	Words
#1	battledore, audience, net, wall, lamp, window, floor, stuff, athlete
#2	oar, rowboat
#3	sailingboat, sky, building, house, water
#4	rope
#5	spectator, player
#6	mallet, wicket, croquet, ball, tree, grass, plant
#7	horse
#8	skier, ski

(a) (b) (c)

Figure 5. Co-clustering results on UIUC-Sport. The table above shows word clusters. (a-c) Example of image descriptors (red circles) correlated to the word cluster #7 ("horse") are overlaid on images.

data). Each image has 11 textual words on average. We remove ones that occur less than 3 times, and obtain a vocabulary of 186 unique words ($W = 186$).

For both datasets, we randomly sample 50000 image descriptors (SIFT) extracted from the training data, and use them for visual dictionary learning. For coding, we employ ML coding (see 3.2) for CD-BCC and vector quantization for K-means and co-clustering methods. We use kNN classifiers (k=5) based on the similarity of image representation vectors encoded by the learned dictionary.

The results are shown in Table 1. First, our approach outperforms all the other methods with statistical significance $p < 0.05$. This may be because that the unified learning of the intermediate descriptor distributions and their correlation to the textual words allows the visual dictionary to capture both visual and textual properties of the training images. Second, most of the multimodal learning methods outperform K-means. This result suggests that text

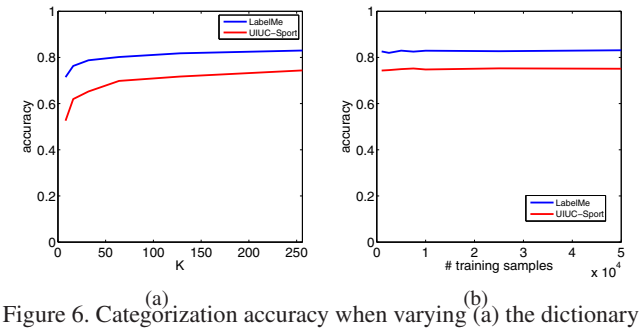


Figure 6. Categorization accuracy when varying (a) the dictionary size K where 50000 training samples are used and (b) the number of training samples in which we fix $K = 256$.

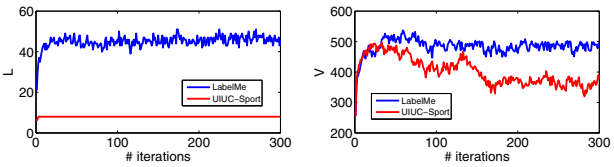


Figure 7. Numbers of (a) word clusters L and (b) descriptor distributions V at each iteration step.

information significantly improves the performance of visual dictionaries for image categorization. Third, most co-clustering methods show higher performance than the state-of-the-art Bayesian dictionary learning [14]. This suggests that co-clustering can be more promising than multimodal topic modeling for multimodal visual dictionary learning. Note that [24] compared their model to two other Bayesian topic models, CorrLDA [2] and a hierarchical Bayesian model [6], and consistently outperformed these two methods. Therefore our CD-BCC is also better than these two models in the same setting.

We also report confusion matrices in Fig. 4. Our approach successfully identifies most categories correctly. On UIUC-Sport, we found that "bocce" is still difficult to distinguish from others. However, ours better distinguishes some categories than the method in [14] (e.g., "polo" from "croquet" and "badminton"). On LabelMe, some category pairs like "coast" and "open country" are commonly confused by both methods. However, our approach better distinguishes "street" from "inside city" and "highway" from "open country". One reason can be that CD-BCC successfully discovers image descriptor clusters related to a specific image category via co-clustered textual words. For instance, as shown in Fig. 5, our CD-BCC discovered a word cluster related to "horse" (#7), which is clearly relevant to "polo" but irrelevant to "croquet" and "badminton". Fig. 5(a-c) show that image descriptors correlated to the "horse" word cluster are actually extracted from the parts of horses.

We also analyze the performance when varying dictionary size K and the number of training samples (image descriptors). The results are shown in Fig. 6 (a) and (b), respectively. Similar to the most existing visual dictionary learning methods, the performance is somewhat sensitive to K. On the other hand, surprisingly, the performance is not

Table 2. Categorization accuracy on Caltech101 dataset. Top two scores in each setting are highlighted.

	# train=5	10	15	20	25	30
ScSPM [27]	-	-	**67.0**	-	-	**73.2**
K-SVD [1]	49.8	59.8	65.2	**68.7**	**71.0**	**73.2**
LLC [25]	**51.2**	59.8	65.4	67.7	70.2	**73.4**
SCC [3]	**51.2**	**60.4**	65.2	67.7	69.5	72.8
ITCC [4]	51.1	58.8	65.8	67.5	70.2	73.0
NMTF [5]	49.9	55.3	59.6	62.8	65.3	69.3
BCC [22]	50.2	60.1	65.2	68.1	70.3	**73.2**
Ours	**51.5**	**60.5**	**66.2**	**68.6**	**70.4**	**73.2**

significantly affected by the number of training samples in spite of the fact that we have randomly chosen training samples. The major reason can be that our CD-BCC trains a visual dictionary based on a statistical relationship between distributions of image descriptors and textual words, which can be stable (robust) against the number of training samples as well as ways to choose them.

Fig. 7 shows estimated V (the number of descriptor distributions) and L (the number of clusters for text words) at each Gibbs sampler iteration step. For both datasets, V and L are converged after about 200 iterations.

4.1.2 Evaluation on Caltech101 Dataset

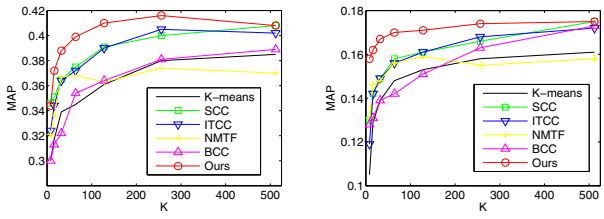
We next analyze the performance on Caltech101 dataset. We select this dataset because this is frequently used to evaluate the performance of visual dictionary learning methods. We follow the common experimental settings for this dataset, i.e., we use 5, 10, 15, 20, 25 and 30 images per category for training visual dictionaries and classifiers, and test on the rest. Caltech101 originally does not include any textual words, we thus directly use the class labels as textual words. We employ sparse coding with SPM [13] and linear SVM to perform image categorization based on the visual dictionary trained by our CD-BCC. We compare ours to ScSPM [27] (sparse coding and SPM with K-means dictionary + linear SVM), two visual dictionary learning methods (K-SVD [1] and LLC [25]), and four co-clustering based methods (SCC, ITTC, NMTF, and BCC). We fix $K = 1024$.

The results are shown in Table 2. Our CD-BCC is highly competitive to the other methods and shows the best performance when the number of training images is small. Note that our method is also comparable to some other recent supervised visual dictionary learning methods like [12, 30]. The performance may be further improved by combining our visual dictionary with more sophisticated coding approaches like LLC.

4.2. CBIR

We analyze CBIR performance of our approach. We used the following FLICKR and WIKIPEDIA datasets:

FLICKR: This dataset[5] has originally been developed based on Pascal VOC 2007 Flickr images for semi-supervised multimodal learning [7]. We use this dataset because this contains images with tags and manually annotated class labels, well fits to our scenario. Total 9963 images of 20 categories of objects (multiply labeled) are divided into training/testing sets by the publisher: the training set (for visual dictionary learning) consists of 5011 images and 3095 out of them are associated with at least one tag from 804 unique tags ($W = 804$), where the testing set has 4952 images.



(a) (b)

Figure 8. CBIR performance (MAP) on (a) FLICKR dataset and (b) WIKIPEDIA dataset. Results on various dictionary sizes $K \in \{8, 16, 32, 64, 128, 256, 512\}$ are reported.

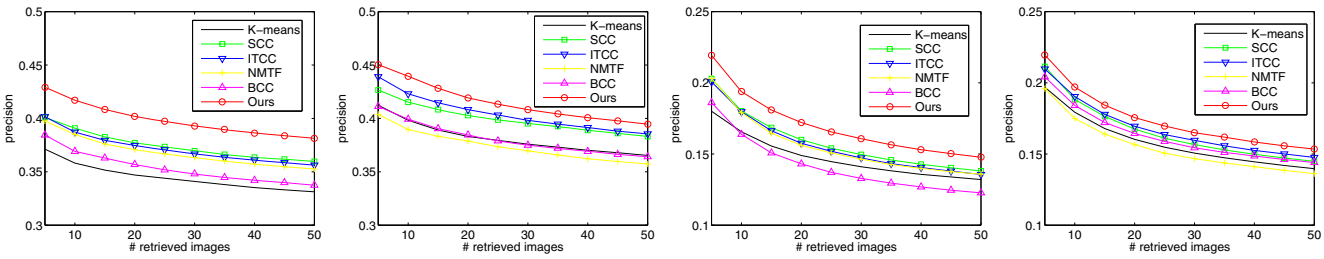
WIKIPEDIA: This dataset consists of Featured Articles in Wikipedia[6], and first used in [21] for image-text retrieval task. We select this dataset because this contains Wikipedia images, texts, and class labels just like FLICKR dataset. Total 3482 articles consists of 21672 images were collected. Each article is categorized into one out of 30 classes defined by Wikipedia. We evenly and randomly split each class to construct training and testing datasets. For textual words, we first extracted only noun terms from all the documents, and selected 300 most frequent words ($W = 300$).

We choose one from the test dataset as a query image, and try to retrieve images from the rest in the same category with the query. We use histogram intersection to perform retrieval, and use mean average precision (MAP) to measure the performance. We compare our method to K-means and four existing co-clustering methods (with BoVW representation). We employ ML coding for this experiment.

Fig. 8 shows the results. Except for the case of $K = 8$ on FLICKR, our approach shows the best performance. In most cases, co-clustering approaches achieve higher MAP than K-means. These results suggest that (i) leveraging textual words via co-clustering is effective for CBIR, and (ii) our CD-BCC is the promising co-clustering approach for this purpose. Fig. 9 shows precisions under different number of retrieved images. Our approach consistently outperforms all the other methods.

5. Conclusion

Focusing on the scenario where images are associated with textual words, we presented a Bayesian approach to multimodal visual dictionary learning. We proposed a novel Bayesian co-clustering, CD-BCC, to learn a single visual dictionary based on the distributions of image descriptors over the continuous space, as well as the relationship between image descriptors and textual words. Extensive ex-



(a) (b) (c) (d)

Figure 9. Precision at different numbers of retrieved images: (a) $K = 64$ and (b) 256 on FLICKR, and (c) $K = 64$ and (d) 256 on WIKIPEDIA.

periments validated values of textual words in improving visual dictionary learning, where our model showed superior performance over several recent methods. Our future work will focus on exploring faster alternative inference algorithms like variational method or slice sampling.

A. Sampling Distribution

The sampling distributions for $\boldsymbol{\omega}$, $\boldsymbol{z}^x$, and $\boldsymbol{z}^w$ are:

$$
\begin{aligned}
&p(\omega_i = t|\mathbf{X}, \mathbf{R}, \boldsymbol{\omega}_{-i}, \boldsymbol{z}^x, \boldsymbol{z}^w) \\
&\propto (m_{t,-i} + \gamma/V) \\
&\times \left\{ \frac{|\Sigma_t|^{1/2}}{|\Sigma_x|^{m_t/2}} \exp\left[-\frac{1}{2} \text{trace}\left(\Sigma_x^{-1} S_t - \Sigma_t \hat{\mu}_t \hat{\mu}_t^T \right) \right] \right\} \\
&\times \left\{ \prod_{k,l}^{K \times L} R^{k,l} \Gamma(m_{k,l} + \beta) \left(\frac{\phi}{n_{k,l}\phi + 1} \right)^{(m_{k,l}+\beta)} \right\} \quad (7)
\end{aligned}
$$

$$
\begin{aligned}
&p(z_v^x = a|\mathbf{X}, \mathbf{R}, \boldsymbol{\omega}, \boldsymbol{z}_{-v}^x, \boldsymbol{z}^w) \\
&\propto \left(m_{a,-v}^x + \zeta/K \right) \\
&\times \left\{ \prod_{l}^{L} \Gamma(m_{a,l} + \beta) \left(\frac{\phi}{n_{a,l}\phi + 1} \right)^{(m_{a,l}+\beta)} \right\} \quad (8)
\end{aligned}
$$

$$
\begin{aligned}
&p(z_j^w = b|\mathbf{X}, \mathbf{R}, \boldsymbol{\omega}, \boldsymbol{z}^x, \boldsymbol{z}_{-j}^w) \\
&\propto \left(m_{b,-j}^w + \eta/L \right) \\
&\times \left\{ \prod_{k}^{K} \Gamma(m_{k,b} + \beta) \left(\frac{\phi}{n_{k,b}\phi + 1} \right)^{(m_{k,b}+\beta)} \right\} \quad (9)
\end{aligned}
$$

where, $m_t/m_{t,-i}$ is the number of image descriptors in t-th descriptor distribution with/without i-th image descriptor. $m_{k,l}$ and $n_{k,l}$ are the sum of $r_{v,j}$ and the number of elements in (k,l)-th block. $m_{k,-v}^x$ ($m_{l,-j}^w$) is the number of descriptor distributions (textual words) in k-th (l-th) cluster without v-th descriptor distribution (j-th textual word). $\hat{\mu}_t \equiv \Sigma_x^{-1} \sum_{i \text{ s.t. } \omega_i = t} x_i$, $S_t \equiv \sum_{i \text{ s.t. } \omega_i = t} x_i x_i^\top$, $\Sigma_t^{-1} \equiv \Sigma_0^{-1} + m_t \Sigma_x^{-1}$, and $R^{k,l} \equiv \prod_{v,j \text{ s.t. } z_v^x = k, z_j^w = l} \frac{1}{r_{v,j}!}$. Γ is Gamma function. The non-parametric version of sampling distributions (discussed in Section 3.3) can also be derived by replacing the first parentheses in Eq. (7) and Eq. (9) by Eq. (5) and Eq. (6) respectively. For the detail of the derivations, see the supplementary material.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionries for sparse representation. *IEEE Trans. on Signal Processing*, 54:4311–4322, 2006. 2, 7

[2] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, 2003. 3, 6

[3] I. Dhillon. Coclustering documents and words using bipartite spectral graph partitioning. In *KDD*, 2001. 2, 5, 7

[4] I. Dhillon, S. Mallela, and D. Modha. Information-theoretic co-clustering. In *KDD*, 2003. 2, 5, 7

[5] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *KDD*, 2006. 2, 5, 7

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 6

[7] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010. 1, 2, 7

[8] Y. Jia, M. Saizmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 3

[9] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*, 2011. 1, 2

[10] Z. Jiang, G. Zhang, and L. Davis. Submodular dictionary learning for sparse coding. In *CVPR*, 2012. 1, 2

[11] D. Kim, M. Hughes, and E. Sudderth. The nonparametric metadata dependent relational model. In *ICML*, 2012. 2

[12] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *TPAMI*, 31:1294–1309, 2009. 1, 2, 7

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 7

[14] L. Li, M. Zhou, G. Sapiro, and L. Carin. On the integration of topic modeling and dictionary learning. In *ICML*, 2011. 3, 5, 6

[15] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 3

[16] X.-C. Lian, Z. Li, C. Wang, B.-L. Lu, and L. Zhang. Probabilistic models for supervised dictionary learning. In *CVPR*, 2010. 2

[17] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. In *TPAMI*, 2012. 1, 2

[18] E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis. Modeling dyadic data with binary latent factors. In *NIPS*, 2006. 2

[19] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regressionmulti-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010. 3

[20] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR*, 2007. 2

[21] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, 2010. 3, 7

[22] H. Shan and A. Banerjee. Bayesian co-clustering. In *ICDM*, 2008. 2, 5, 7

[23] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *J. American Stat. Assoc.*, 101:1566–1581, 2006. 5

[24] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009. 3, 5, 6

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 1, 2, 4, 7

[26] P. Wang, K. Laskey, C. Domeniconi, and M. Jordan. Nonparametric bayesian co-clustering ensembles. In *SDM*, 2011. 2

[27] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 4, 7

[28] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011. 2

[29] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the socialweb. In *ACL*, 2009. 3

[30] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, 2010. 1, 2, 7

[31] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, 2012. 2

[32] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011. 3