Correlated Topic Vector for Scene Classification

Pengxu Wei, Fei Qin, *Member, IEEE*, Fang Wan, Yi Zhu, Jianbin Jiao, *Member, IEEE*, and Qixiang Ye, *Senior Member, IEEE*

***Abstract*— Scene images usually involve semantic correlations, particularly when considering large-scale image data sets. This paper proposes a novel generative image representation, correlated topic vector, to model such semantic correlations. Oriented from the correlated topic model, correlated topic vector intends to naturally utilize the correlations among topics, which are seldom considered in the conventional feature encoding, e.g., Fisher vector, but do exist in scene images. It is expected that the involvement of correlations can increase the discriminative capability of the learned generative model and consequently improve the recognition accuracy. Incorporated with the Fisher kernel method, correlated topic vector inherits the advantages of Fisher vector. The contributions to the topics of visual words have been further employed by incorporating the Fisher kernel framework to indicate the differences among scenes. Combined with the deep convolutional neural network (CNN) features and Gibbs sampling solution, correlated topic vector shows great potential when processing large-scale and complex scene image data sets. Experiments on two scene image data sets demonstrate that correlated topic vector improves significantly the deep CNN features, and outperforms existing Fisher kernel-based features.**

***Index Terms*— Correlated topic vector, Fisher kernel, generative feature learning, semantic correlation.**

I. Introduction

SCENE classification has been widely explored, promoting related computer vision tasks including object recognition [1], [2], image retrieval [3]–[5], and intelligent robot navigation [6], [7]. A scene image is usually composed of several semantic entities e.g., *sky*, *rock*, *street*, and *car*. These entities are often organized in unpredictable layouts [8], [9] and shared with multiple categories, which invite intra-class variability and inter-class similarity for scene recognition. Scene labels, e.g., *coast*, *village*, *coast*, and *inside city*, are equivalently the overall cognition and high-level abstract of scene images, which are difficult to be captured using low-level visual features. These factors make scene image recognition much more challenging than object-centric image classification.

The conventional visual recognition method extracts local visual descriptors and encodes them into a global representation of one image. Many efforts on this strategy for scene recognition focus on two problems: (1) how to characterize semantics, commonly known as topics or themes, explicitly or implicitly, and (2) how to encode superior scene representation based on the semantics. The first class of semantics consists of object-centric approaches that model pre-defined explicit semantics or scene categories. It annotates regions with corresponding explicit themes and trains specific theme classifiers. One popular strategy of theme labeling leverages a group of object detectors pre-trained on available object-centric image datasets [10]. The other one utilizes given scene categories and assumes that a specific scene category is shared for all the patches of one image [11]. These approaches rely on theme performance heavily since they attempt to independently discover potential themes. The second class of semantics devotes to scene-centric representation [12]–[15]. It is learned from an entire image and generates a holistic description with the aid of implicit themes. And it works without explicit image segmentations, manual theme annotations or extensive object detections.

The scene-centric representation is conventionally built on Bag-of-Words (BoW) that encodes an image as an orderless collection of local descriptors. BoW takes cluster centers resulted from k-means as semantics and encodes semantic histograms as features. Without any doubt, the lossy BoW quantization procession of local descriptors is bound to induce word ambiguity [16] including synonymy (different visual words may represent the same semantic) and polysemy (the same visual word may represent different semantics in different contexts). As shown in the first row of Fig. 1, BoW features present significant differences between the first two images even though both images belong to the *village* scene, which indicates its limited capacity for the intra-class variance. Generative models from statistical text literature, e.g., probabilistic Latent Semantic Analysis (pLSA) [17] and Latent Dirichlet Allocation (LDA) [18], improve BoW by dealing with the ambiguity problem [16], and introduce intermediate latent topic features that are scene-centric [12]–[14]. In the third row of Fig. 1, it is observed that a group of themes, *sky-rock-house-tree*, generally co-exist in the *village* scene. Obviously, a scene exhibits a strong semantic/theme correlation property, and more importantly, this property is specific to a scene category distinguishing itself from others. Unfortunately, such correlation is ignored in most existing

Manuscript received February 24, 2016; revised December 19, 2016 and March 18, 2017; accepted March 29, 2017. Date of publication April 13, 2017; date of current version May 9, 2017. This work was supported in part by the National Nature Science Foundation of China under Grant 61401426 and Grant 61671427, in part by the Beijing Municipal Science and Technology Commission under Grant Z161100001616005, and in part by the Science and Technology Innovation Foundation of Chinese Academy of Sciences under Grant CXJJ-16Q218. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paul Rodriguez. *(Corresponding author: Jianbin Jiao.)*

The authors are with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: weipengxu11@mails.ucas.ac.cn; fqin1982@ucas.ac.cn; wanfang13@mails.ucas.ac.cn; zhuyi215@mails.ucas.ac.cn; jiaojb@ucas.ac.cn; qxye@ucas.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2694320

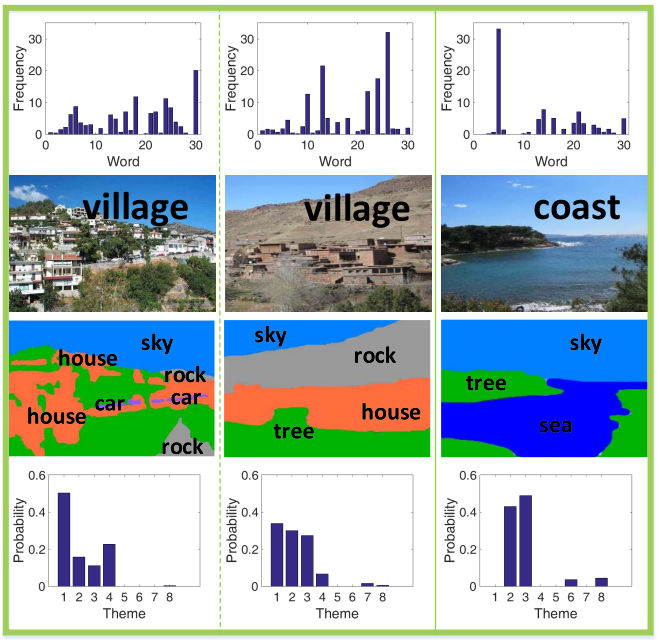


Fig. 1. Examples of *village* and *coast* scene images. In the first row, histograms of visual words are shown. Themes are provided for three images, among which two images of scene *village* and one image of scene *coast* present in the second row. Their corresponding theme probability distributions are shown in the fourth row.

work besides BoW. For example, LDA imposes Dirichlet distribution prior on the topic proportions [18], which poorly assumes that themes/topics are independent of each other.

In this paper, we propose a new feature representation, named Correlated Topic Vector (CTV), which utilizes the Correlated Topic Model (CTM) [19], [20] to capture the correlations between themes as a latent semantic representation. CTM replaces the Dirichlet distribution prior of the classical Latent Dirichlet Allocation model with a more flexible logistic normal distribution [21] that incorporates a covariance measurement among topics. This makes it possible to describe more realistically the fact that the presence of one latent topic may be correlated with the presence of another. However, the latent semantic representation derived from CTM with the conventional way that just considers the latent topic distributions [13], [14], fails to perform well consistently. This similar case happens to other topic models [22], e.g., pLSA and LDA, the latent topic representations of which are generally believed with limited discriminative capability due to the unsupervised learning.

The proposed CTV further explores the contributions of low-level visual words to the learning of middle-level topics from the information geometry view in essence. This is different from BoW and latent semantic representations since BoW depends on visual word co-occurrence counts and latent semantic representations focus on topic distributions. For two images from different categories, regions with similar appearances tend to follow the same visual words and limited topics hold insignificant differences for recognition tasks. Built on Fisher Kernel [23], the CTV takes these properties into account, combining the benefits of generative and discriminative approaches.

To demonstrate the up to date performance, the proposed CTV is implemented on Convolutional Neural Network (CNN) [24] features. For scene recognition, it is demonstrated that the features extracted from a fully connected layer of CNN trained on ImageNet [25] show a clear semantic clustering, and the latter layers learn semantic features [26]. Therefore, it can be utilized as an alternative representation without any object detection or segmentation efforts. It is an intuition that regarding CNN feature of the hidden fully connected layer as a learned soft-assignment word histogram, avoiding to build a vocabulary relying on CNN as local descriptors.

To summarize, this work has the following contributions:

1. Rooted in the classical Correlated Topic Model (CTM), we propose a new image descriptor, Corrected Topic Vector (CTV), targeting at modeling and leveraging the topic correlations for scene image classification.

2. For the first time, we derive the formal expression of CTV in the Fisher Kernel space, making it theoretically possible to use Fisher Kernel to enhance the discriminative capacity of CTV.

3. We provide an efficient Gibbs sampling solution and make it feasible to train CTM and extract CTV on large-scale datasets.

In the remainder of the paper, we review related work in Section II and discuss the details of correlated topic vector in Section III. Experimental results are provided in Section IV. We conclude in Section V.

II. Related Work

Inspired from text categorization [27], BoW [28] has been widely used for image recognition. It characterizes an image with visual word co-occurrence. Hard word assignments and histogram encoding induce the loss of image spatial information and the semantic ambiguity for each word, let alone semantic correlation that is a noticeable attribute for scene recognition. The intermediate "theme" or "semantic" representation for scene images is an extension of BoW and attempts to fill the semantic gap between the low-level image features and the high-level semantic concepts.

Exploiting explicit themes assigned directly to patches or regions suffers from theme annotation efforts or unreliable detection results of diverse objects. Li *et al.* [10], [29] propose "Object Bank" (OB) that deploys a large number of object detectors at multiple scales to obtain the probability of objects appearing at each pixel. It detects 177 categories of objects at 12 scales and 21 spatial pyramid grids. But it is hard to generalize OB to large-scale scene image sets such as SUN 397 [30] or Places 205 [23], duo to extensive detections. Besides, Li *et al.* manually illustrate the identities and semantic relations among 177 objects carefully selected from 1000 objects; however, these relations are not employed for scene recognition.

Some works are devoted to Fisher Kernel [31] to improve BoW. Aitchison [21] provide a formulation of Fisher Kernel for classification tasks. Perronnin and Dance introduce Fisher Kernel derived from Gaussian Mixture Models (GMM) for

the image representation. The resulted Fisher vectors benefit from powerful local feature descriptors [32]. The Vector of Locally Aggregated Descriptors (VLAD) [33] improves BoW to produce a compact representation. Fisher kernels have already been applied to the problem of image categorization built on generative models [34]. Dirichlet-based GMM Fisher Kernel [35] is applied as a way of feature transformation for image classification, assuming that L_1-normalized histogram-based local descriptors could be modeled by the Dirichlet distribution.

CNN features have recently achieved spectacular results on the ImageNet object recognition challenge. Their success has encouraged the community to use CNN feature embedding for scene classification to replace the conventional SIFT-FV architecture. For example, Gong *et al.* represent a scene image as a collection of fully connected layer activations extracted from local patches and build VLAD embedding for image recognition. Dixit *et al.* incorporate semantics into the Fisher Kernel framework. They extract CNN features of local patches and consider them as Semantic MultiNomial (SMN) descriptors. When local semantic descriptors are modeled as a multinomial distribution, with the help of Dirichlet Mixture Models (DMM), the DMM FV is induced as a more natural embedding than the GMM FV. Besides, the natural parameterization transformation alleviates highly non-Euclidean property of SMN descriptors. A semantic FV is then computed as a GMM FV in the space of the natural parameters.

A considerable number of works built on the Fisher Kernel framework for image recognition have made great strides, but they are generally assumed that patches of all the images are independently and identically drawn from the involved generative models. Obviously, the independent and identically distributed (i.i.d.) assumption violates intrinsic image characteristics. In addition, semantic correlation is seldom considered in existing works. Considering an image as an unordered set of regions, Cinbis *et al.* [36], [37] utilize the Dirichlet prior distribution to parameterize the variables varying across images. They consider models, e.g., Latent Dirichlet Allocation and latent Gaussian Mixture Models, which capture the dependencies among local image regions. For latent GMM, they treat the parameters of GMM as latent variables with prior distributions learned from data, and apply the Fisher Kernel principle by taking the gradient of the log-likelihood of the observed data with respect to the hyper-parameters. These hyper-parameters control the priors on the latent model parameters. Despite the wide exploration of latent semantics in these works, the semantic correlation remains not considered.

III. Methodology

Based on the hypothesis that CTM could reasonably model the relationship among topics for latent semantic features and Fisher Kernel can further enhance the discriminative capacity, the task of scene classification will be pretty straightforward: firstly estimate the parameters of CTM from a training set, and then build the Correlated Topic Vector with the aid of Fisher Kernel framework for both the training and test images. The CTV will be utilized as the final feature representation, which can be fed to a linear SVM classifier to recognize different scene categories. In this section, the detailed derivation and solutions of CTV will be discussed. We firstly introduce latent semantics, by which semantic co-occurrence implies certain correlations. We then construct the CTV by utilizing both Variational Bayesian (VB) method and Gibbs Sampling (GS) method. The basic scheme of CTV encoding has been shown in Fig. 2.

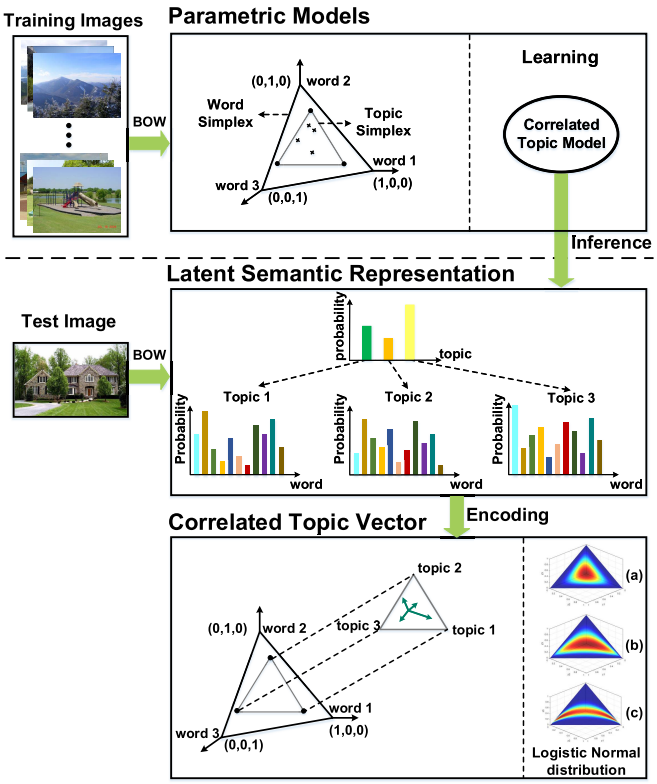


Fig. 2. Correlated Topic Model learning and Correlated Topic Vector encoding.

TABLE I
GENERATIVE PROCESS OF THE TOPIC MODEL

1. Draw $\eta_d \vert \{\mu, \Sigma\} \sim \mathcal{N}\{\mu, \Sigma\}$, where μ and Σ are parameters, mean and covariance.
2. For each word $n \in \{1, ..., N_d\}$: (a) Draw the topic assignment of z_n from $Mult(f(\eta_d))$, where $f(\eta_d)$ is a natural parameterization of the topic proportions η_d to the mean parameterization θ_d; (b) For each topic, draw word $w_{d,n} \vert \{z_n, \beta\}$ from $Mult(\beta_{z_n})$.

A. Latent Semantic Representation

CTM is introduced as a generative model for scene image data. The motivation is two folds: firstly to remove the independence assumptions implicitly of the Dirichlet distribution on topic proportions [36], [37], and secondly to further model the correlation structures among topics by a logistic normal prior [39]. The generative process of the CTM has been stated in Table I.

Given a dataset that consists of D images, each image is represented as a collection of visual words from a

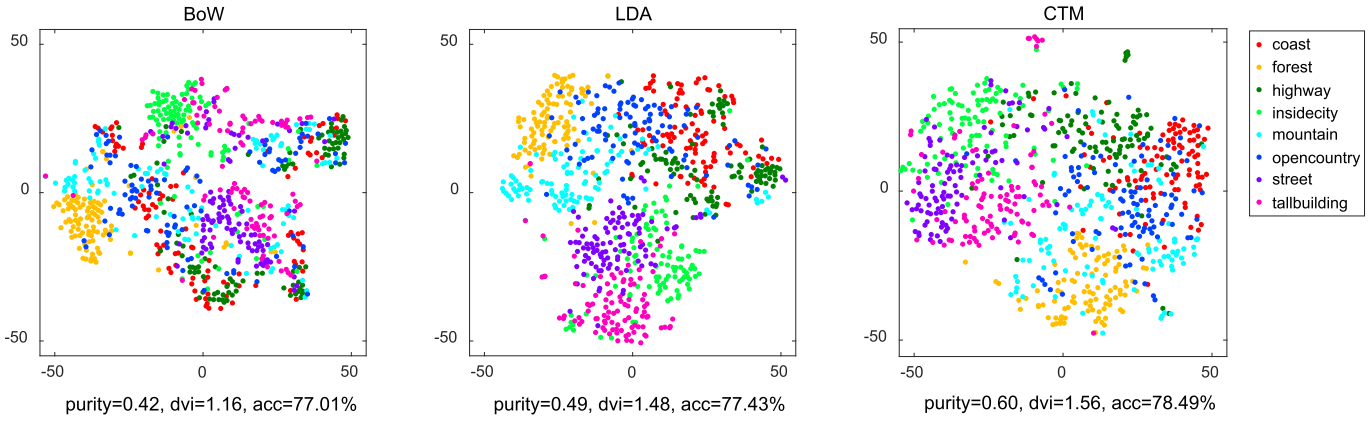


Fig. 3. Feature visualization using t-sne [38]. (Best viewed in color.)

vocabulary containing V visual terms. Formally, let $w_d = \{w_{d,n}, n \in 1, ..., N_d\}$ denote the visual word indices corresponding to N_d patches sampled in an image d, where $w_{d,n}$ is the word assignment for its n-th patch. In CTM, an image is modeled as mixtures over K latent topics, where each topic is represented by a multinomial distribution over V visual words. Specifically, given a certain topic, each word is sampled with respect to a multinomial distribution and its probability is parameterized by a matrix $\beta = (\beta_{ij})_{K \times V}$.

The essential of CTM is a more flexible logistic normal distribution [21], which has been employed to model the realistically latent topic structure. As discussed in Section I, this is hinted at the fact that one topic may be correlated with others. Since CTM is based on the logistic normal distribution, such correlation among topics could be reasonably modeled by incorporating the covariance structure [20]. The logistic normal distribution, parameterized by K dimensional mean vector μ and $K \times K$ covariance matrix Σ, both of which are hyper-parameters, is then imposed on topic proportions as a prior in CTM. The topic proportion of image d is termed as $\theta_d = [\theta_d^1, ..., \theta_d^i, ..., \theta_d^K]$, where

$$\theta_d^i = f(\eta_d^i) = \exp \eta_d^i / \sum_{i'} \exp \eta_d^{i'}, \tag{1}$$

and i or i' indicates the i-th or i'-th topic of K topics. It assumes that η_d is subject to a normal distribution $\mathcal{N}\{\mu, \Sigma\}$. Consequently, $f(\eta_d)$ maps η_d to its mean parameterization θ_d located as a point on the $K-1$ topic simplex. To highlight, the parameter Σ interprets the relationship among topics.

As shown in Fig. 2, the topics are shared by all images in the dataset. But the topic proportions, i.e., θ_d, definitely vary stochastically across images, as they are randomly drawn from the prior distribution. After θ_d are obtained, words could be drawn from each topic in the collection according to β.

It is very straightforward to make the hypothesis that the topic proportions θ for each image could be utilized as the desired latent semantic representation. Two main reasons are: (1) topic proportions θ remain the image-specific property; (2) θ imply correlation-ship among topics stemming from a logistic normal prior. To demonstrate the performance of CTM, in Fig. 3, we visualize three types features on the SCENE 8 dataset [40]: BoW, latent semantic representations of LDA and CTM. Three measures, including the cluster purity [41], Dunn Validity Index (dvi) [42], and the average classification accuracy (acc) [43], are utilized to evaluate these features. Cluster purity denotes the ratio between the dominant class in the cluster and the size of the cluster. The larger the purity is, the better the clustering solution is [41]. The dvi identifies how compact the clusters are [42]. It can be observed that latent semantic features derived from CTM demonstrate a superior cluster effect for all the scene categories in the semantic feature space, as well as a better classification performance.

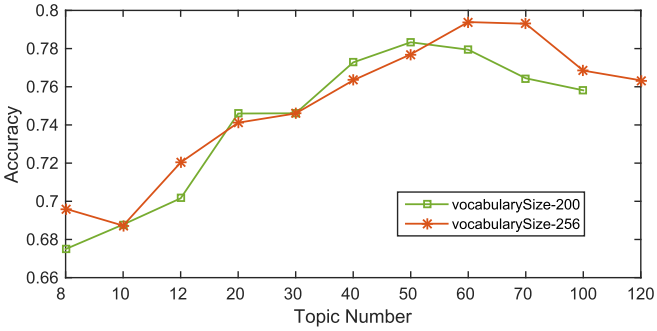


Fig. 4. Performance of CTM based latent semantic representation on the SCENE 8 dataset.

It is noted that the performance of CTM based latent semantic representation can not consistently increase with the increasing number of topics yet. Experiments shown in Fig. 4 validate this situation. Given 200 and 256 vocabulary sizes, the topic representations perform respectively better and better with the topic number increasing, but the accuracy decreases when the topic number is larger than 50 and 60, respectively. A similar situation of LDA has been also demonstrated in [22]. This limitation of latent semantic representations usually does not result from the poor statistical estimation, but from the intrinsic ambiguity of the underlying BoW representation [15]. Another reason is that latent semantic features stemming from word co-occurrence fail to utilize the statistical information between semantics/topics and words. These two reasons explain why latent features derived from CTM should not be simply utilized, although it better characterizes scene semantics through the modeling of the correlation structure among topics.

To further encode informative features based on semantics, we attempt to explore the contributions of low-level words to the learning of middle-level semantics from the information geometry view. To this end, we propose the feature encoding scheme of Correlated Topic Vector aided by the Fisher Kernel Framework [23], which can integrate the benefits of generative and discriminative approaches.

B. Correlated Topic Vector

We first discuss the derivation of a formal expression of CTV features for an image by utilizing the Variational Bayesian method. According to the log-likelihood, to derive the CTV, we mainly compute the Fisher score and the Fisher information matrix with respect to the model parameters $\Theta = \{\mu, \Sigma, \beta\}$, i.e., hyper-parameters $\{\mu, \Sigma\}$ and global parameter β.

The log-probability of an image d is $L = \log p(w_d|\mu, \Sigma, \beta)$ defined by:

$$p(w_d|\mu, \Sigma, \beta) = \int p(\eta|\mu, \Sigma)\Big(\prod_{n=1}^{N_d}\sum_{z_n} p(z_n|\eta)p(w_{d,n}|z_n, \beta)\Big)d\eta, \quad (2)$$

where z_n denotes a vector of topic assignments of word $w_{d,n}$ with only one component equivalent to 1 and others to 0, i.e., the occurrence of word n in image d.

It is obvious that the logistic normal prior distribution of topic proportions $p(\eta|\mu, \Sigma)$ is non-conjugate to the multinomial posterior distribution of topic assignments $p(z_n|\eta)$ [20]. As a result, it is hard to analytically compute the integrals in Equation (2). In other words, we cannot directly derive the gradient of the log-likelihood to obtain CTV features.

We resort to the Variational Bayesian method [44] to derive the formal expression. Variational Bayesian is an approximate approach that optimizes a deterministic objective lower bounded on the data log-likelihood [44]. With mean-field assumptions [19], the original graphic model is simplified with variational parameters $\{\lambda, \nu^2, \phi\}$. In this case, $L = L_{VB} + D_{KL}(q||p) \geq L_{VB}$, where D_{KL} is the Kullback-Leibler (KL) divergence between distribution q and p. L_{VB} denotes the lower bound of log-likelihood. L can be approximated as L_{VB}:

$$L_{VB} = E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N_d} E_q[\log p(z_n|\eta)] + \sum_{n=1}^{N_d} E_q[\log p(w_{d,n}|z_n, \beta)] + H(q), \quad (3)$$

where $E_q[\cdot]$ is the expectation with respect to the variational distribution q whose parameters are $\{\lambda, \nu^2, \phi\}$, and $H(q)$ denotes the entropy of this distribution. Variational parameters $\{\lambda, \nu^2, \phi\}$ are K-dimension image-specific vectors. Details on how to obtain variational parameters $\{\lambda, \nu^2, \phi\}$ and model parameters $\{\mu, \Sigma, \beta\}$ can be found in [20].

Now, we derive the formal expression of the CTV for image d based on the learned model parameters $\Theta = \{\mu, \Sigma, \beta\}$. Its form is $\varphi_{[\Theta]} = {I_{[\Theta]}}^{-1/2} u_{[\Theta]}$. $u_{[\Theta]} = \partial L/\partial\Theta$ denotes the Fisher score which is the partial derivative of the log-likelihood, and $I_{[\Theta]} = E[{u_{[\Theta]}}^T u_{[\Theta]}]$ is the Fisher information matrix. $u_{[\Theta]}$ represents the velocities passing through Θ along the coordinate curves, while $I_{[\Theta]}$ plays the role of a metric tensor. Under certain regularity conditions, the Fisher information matrix is the negative of the expectation of the second derivative with respect to Θ. $I_{[\Theta]} = E[{u_{[\Theta]}}^T u_{[\Theta]}]$ can be written as $I_{[\Theta]} = -E[\partial^2 L/\partial\Theta^2]$.

The Fisher scores based on hyper-parameters $\{\mu, \Sigma\}$ are

$$u_{[\mu]} = \partial L/\partial\mu = \Sigma^{-1}(\lambda_d - \mu), \quad (4)$$

$$u_{[\Sigma^{-1}]} = \partial L/\partial\Sigma^{-1} = 1/2(\Sigma - diag(\nu_d^2) - (\lambda_d - \mu)^T(\lambda_d - \mu)). \quad (5)$$

The Fisher score based on global parameter β is $u_{[\beta]} = (u_{[\beta_{ij}]})_{K\times V} = (\partial L/\partial\beta_{ij})_{K\times V}$, and

$$\partial L/\partial\beta_{ij} = \sum_{n=1}^{N_d} \phi_{d,ni} w_{d,n}^j/\beta_{ij}. \quad (6)$$

μ and Σ are parameters of the true multivariate Gaussian distribution, and they are learned from all the images. λ_d and ν_d^2 are fit from a single observed image data w_d. $(\lambda_d - \mu)$ measures differences between the mean value of true prior distribution and its approximated variational distribution. It is similar to the term $\Sigma - diag(\nu_d^2)$, which measures the variance differences. $\phi_{d,ni}$ is a multinomial parameter and denotes how likely a word $w_{d,n}$ occurs given the topic i. $u_{[\beta]}$ can be regarded as the expectation of word occurrence whose possibility $\phi_{d,ni}$ is weighted by the global parameter β. To avoid matrix multiplication, we derive the partial derivative of the log-likelihood on Σ^{-1}, the inverse of Σ in Equation (5). The derivations of Equations (4)-(6) are provided in the Appendix.

Fisher information matrix can be simply expressed as:

$$I_{[\mu]} = -E[\partial^2 L/\partial\mu^2] = \Sigma^{-1}, \quad (7)$$

$$I_{[\Sigma^{-1}]} = -E[\partial^2 L/\partial(\Sigma^{-1})^2], \quad (8)$$

$$\begin{aligned} I_{[\beta]} &= -E[\partial^2 L/\partial\beta_{ij}^2] \\ &= -\sum_{n=1}^{N_d} p(w_{d,n}|\theta_d)\partial^2 L/\partial\beta_{ij}^2 \\ &= -\sum_{n=1}^{N_d} p(w_{d,n}|\theta_d)\sum_{m=1}^{N_d} \phi_{d,mi} w_{d,m}^j/\beta_{ij}^2 \\ &= -\sum_{m=1}^{N_d} \phi_{d,mi} w_{d,m}^j/\beta_{ij}^2. \end{aligned} \quad (9)$$

We have immediately three approximated Fisher Vectors on $\{\mu, \Sigma, \beta\}$, where $\varphi_{[\mu]} = {I_{[\mu]}}^{-1/2} u_{[\mu]}$, $\varphi_{[\Sigma]} = {I_{[\Sigma]}}^{-1/2} u_{[\Sigma]}$, $\varphi_{[\beta]} = {I_{[\beta]}}^{-1/2} u_{[\beta]}$. CTV is then obtained by concatenating and normalizing these three vectors. The normalization can be a power normalization or L_2-normalization [45], [46].

C. Gibbs Sampling Based CTV

The Variational Bayesian method mentioned in Section III-B has a deterministic log-likelihood [20], [47], and can be utilized to derive the formal expression of

the CTV. We denote VB based CTV as CTV-VB. However, in the training stage, the learning of the CTM parameters suffers from the costly computations for the approximate Variational Bayesian methods, due to the essential problem of the non-conjugate priors [47], [48]. This problem is expected to aggregate along with the increasingly large and complex datasets. To alleviating this limitation, we resort to the scalable Gibbs sampling algorithm [47], which is simple to carry out and can be essentially parallelized to suit the large-scale data. The Gibbs sampling based CTV is denoted as CTV-GS in the following discussion.

Gibbs sampling avoids deterministic computations of integral terms by subsequently applying a stochastic transition operator to a randomly drawn latent variable rather than optimizing for a lower bound of the log-likelihood, so that it is hard to directly derive the specific form for CTV with Gibbs sampling. One strategy is to approximate the log-likelihood of Gibbs sampling solution. The intuition comes from the evidence that L_{GS} plus the expectation of D_{KL} equals to L_{VB} [49], $L_{GS} = L_{VB} - E_{q(z_T|w)}\{D_{KL}[q(y|z_T, x)||r(y|z_T, x)]\} \leq L_{VB}$, where x is the observed data, z_T is the outcome of iteratively sampling, $y = z_0, z_1, ..., z_{T-1}$ are a series of state variables for each iteration, and $r(y|z_T, x)$ is a specific approximated distribution of $q(y|x, z_T)$. D_{KL} is the KL divergence between distribution q and r [49]. L_{VB} can be regarded as the upper bound on L_{GS}. Therefore we can approximate the log-likelihood of Gibbs sampling controlled by the expectation of D_{KL}. For the CTV-GS, we try to utilize the benefits of both Variational Bayesian and Gibbs sampling to construct CTV. Specifically, the CTV derivations with the Variational Bayesian approach provide the general expression of the CTV features. For CTV-GS whose encoding still relies on the derived formal expression of CTV-VB, the involved parameters are learned with Gibbs sampling while those of the CTV features are learned by Variation Bayesian method in CTV-VB. Both Variational Bayesian and Gibbs sampling methods are approximations of the log-probability of CTM, which characterizes the same dependence among variables in a hierarchical graphic model. The feasibility of such approximation is validated by the experimental results in Section IV.

D. Implementation

We evaluate the proposed CTV based on local descriptors extracted using CNN [24]. Deep CNN has demonstrated remarkable recognition performance [24], [26], [50] and its activated features of deep layers present an excellent generalization of image representation and powerful semantic clustering results [26]. Especially, the rectified linear unit (ReLU) transformation guarantees all the features from deep layers are non-negative. Therefore, they can be considered as a type of soft-assignment BoW with which we implement the deep-BoW.

The conventional CNN-BoW simply encodes BoW by replacing local descriptors [11], [37], [43], e.g., SIFT, with binarized CNN descriptors aided by a clustering derived vocabulary [11], [14], [37], [43]. The utilization of soft-assignment will not only avoid these costly algorithms but also remain more information. The implementation of deep-BoW consists of three procedures: local feature extraction, feature encoding, and pooling. The pipeline of building deep-BoW for one image is shown in Fig. 5. In detail, an image is first divided into patches and sampled on a dense grid with $P \times P$ pixels and a stride size with S-pixel. For each divided patch, fully connected CNN outputs of the seventh (FC7) layer are extracted as local descriptors. These local CNN features are regarded as the encoded local soft-assignment BoW vectors with real activation values. To meet different model complexity requirements, the BoW vectors with desired dimensions could be obtained with a feature element sampling strategy, e.g., fixed sampling, average sampling, or max sampling, as shown in Fig. 6. Finally, a global soft-assignment deep-BoW is achieved by average pooling of local BoW vectors across patches. To facilitate the learning of CTM whose input values are integers, the soft-assignment based deep-BoW with real values is normalized into the final deep-BoW with integer values. Compared with the conventional CNN-BoW, deep Bow benefits from the rich semantic knowledge of the deep CNN networks and involves only simple linear algebraic operations. Experimental comparison in Fig. 6 clearly shows that the proposed deep-BoW outperforms CNN-BoW which is derived from clustering methods, e.g., GMM and k-means, in all scenarios with different sampling strategies.

The deep-BoW vectors will be fed to the CTM in the training stage. The learning of the CTM is based on the VB or GS method. Given the learned CTM parameters, these vectors will be also utilized to generate CTV features with Equations (4)-(9) for classification in the test stage. We implement CTV-VB and CTV-GS (denoted as CTVs for simplicity), at three scale levels (i.e., different P in small scale,

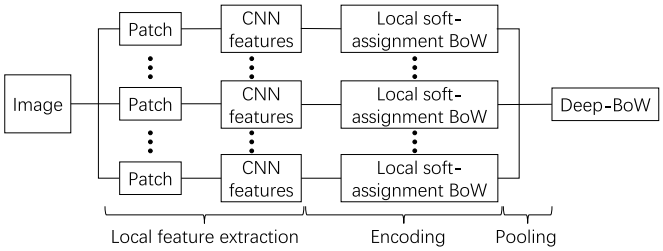


Fig. 5. Pipeline of building deep-BoW.

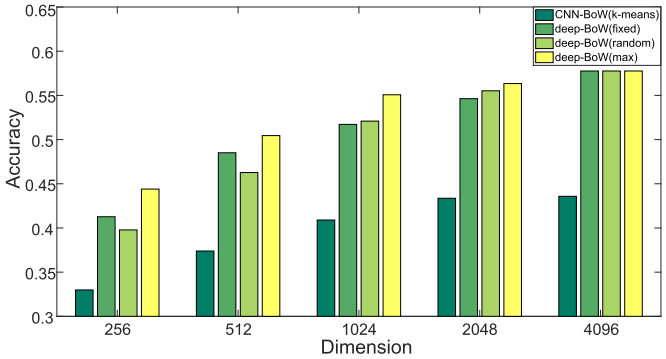


Fig. 6. Comparison of CNN-BoW and deep-BoW on the MIT Indoor 67 dataset.

TABLE II
COMPARISON ON THE SUN 397 DATASET

Methods		Accuracy	Year	Description
Most Relevant Methods	CNN-I [23] (baseline)	42.61	2012	Deep networks trained on the ImageNet dataset
	CTV-GS-I (ours)	**53.21**	–	CTV based on CNN-I with Gibbs sampling, three scale levels
	CTV-VB-I (ours)	**53.35**	–	CTV based on CNN-I with Variational Bayesian, three scale levels
	CNN-P [23] (baseline)	54.32	2014	Deep networks trained on the Places dataset
	CTV-GS-P (ours)	**58.43**	–	CTV based on CNN-P with Gibbs sampling, three scale levels
	CTV-VB-P (ours)	**58.39**	–	CTV based on CNN-P with Variational Bayesian, three scale levels
	DMM FV [55]	49.86	2015	Dirichlet Mixture Model based Fisher Vector; single scale
	Semantic FV [55]	51.80	2015	GMM Fisher Vector with natural parameterization, best three scale levels
	VLAD [43]	51.98	2014	VLAD concatenation of three scale levels
Other State-of-the-art Methods	SPMSM [11]	28.20	2012	Spatial pyramid matching; predefined semantic themes
	Meta-classes [56]	36.80	2014	Classifier-based features
	SUN(MKL) [53]	38.00	2010	Multi-kernel learning
	DeCaF [26]	40.94	2014	DeCaf; global CNN features

middle scale, and large scale). Because of high dimensions of Fisher Vectors, we do not concatenate the three-scale original CTV features to get aggregated multi-scale CTV. Our multi-scale CTV-VB or CTV-GS is derived from the concatenation of SVM scores for three scale features.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed CTV and compare it with the most relevant approaches. Following the experimental settings in [32], [46], and [52], we train one-vs-all linear SVM classifiers on the CTV features and evaluate them with average classification accuracy [43].

A. Setup

1) Datasets: We conduct experiments on two benchmark datasets: SUN 397 [30], [52] and MIT Indoor 67 [53]. SUN 397 is a large-scale dataset for scene recognition. It contains 397 scene categories, and each of ten splits has 50 training and 50 test images per category. MIT Indoor 67 contains 67 scene categories. Images have been split into 5360 training images and 1340 test images, i.e., 80 training and 20 test images per category.

2) Experimental Setup: An image is resized to 256×256 pixels firstly. Three scale levels, which correspond to 256×256, 128×128, 64×64 pixels for the patch sizes, are chosen in this experiment. Patches are sampled with the stride of 32 pixels on all the scale levels. CNN FC7 features are extracted using the Caffe package [50] pre-trained on the ImageNet dataset [25] and the Places dataset [23]. To learn involved parameters for CTV, the deep-BoW with max sampling is fed to the VB based CTM or GS based the scalable CTM [47]. Since Fisher information matrix is immaterial as pointed in [22], we approximate CTV with its Fisher score in the experiments.

In the followings, the deep CNN network trained on the ImageNet dataset [25] is abbreviated to CNN-I, and that trained on Places dataset [23] is abbreviated to CNN-P (PlacesNet). The CTVs learned with the CNN-I are abbreviated to CTV-Is (CTV-VB-I and CTV-GS-I), and CTVs learned with CNN-P are abbreviated to CTV-Ps (CTV-VB-P and CTV-GS-P).

B. Main Results

1) SUN 397 Dataset: Main results on the large-scale SUN 397 have been provided in Table II. We compare the proposed CTVs (CTV-VB and CTV-GS) with the most relevant methods derived from the Fisher Vector framework, and the other state-of-the-art methods. Among these most relevant methods in Table II, DMM FV [54] is single scale, while VLAD [43], Semantic FV [54], CTV-GSs and CTV-VBs are multi-scale, i.e., three scales.

2) ImageNet Based Results: The first group of comparison methods adopts FC7 features from CNN-I as descriptors. These most relevant methods include the baseline CNN and several Fisher Vector based methods: DMM FV [54], Semantic FV [54], and VLAD [43]. In Table II, the baseline CNN-I achieves 42.61% accuracy [23]. The proposed CTV-VB-I achieves 53.35% and CTV-GS-I achieves 53.21% accuracy with accuracy gains up to 10.74% and 10.60%, respectively. Furthermore, compared with other Fisher Vector based methods, CTV-Is also have a better performance. The CTV-I-VB and CTV-I-GS achieve 3.49% and 3.35% accuracy gains over the DMM FV [54], respectively. Although they are Fisher Vector based, one key difference between DMM FV and CTVs is that DMM FV fails to take theme/topic correlations into consideration. The GMM based Semantic FV [54], improves the DMM FV with natural parameterizations of the multinomial parameter vector. Its performance is still 1.55% and 1.41% lower than CTV-VB-I and CTV-GS-I, respectively. Even the concatenation of Semantic FVs at the best four scales achieves 53.0% [54], which is also lower than CTV-Is. Besides, VLAD is pointed as an approximation of Fisher Vector based on GMM [54]. Gong *et al.* [43] report that the CNN-I based multi-scale VLAD can improve the classification performance up to 51.98%. CTV-GS-I and CTV-VB-I also outperform it. In general, among these most relevant methods, DMM FV is built on DMM which assumes that themes/topics are independent of each other; VLAD and Semantic FV, rely on GMM which implies the i.i.d. assumption for all the patches of images [37] without topic correlations. These comparison results with CTV-Is validate the i.i.d. assumption is not always proper to model topics/semantics and that the introduction of topic/semantic correlations improves the derived Fisher Vector based features for the recognition task.

TABLE III
EVALUATION OF FEATURES EXTRACTED AT DIFFERENT SCALES

Methods	MIT Indoor 67				SUN 397			
	256×256	128×128	64×64	Multi-scale	256×256	128×128	64×64	Multi-scale
VLAD [43]	53.73	65.52	62.24	68.88	39.57	45.34	40.21	51.98
Semantic FV [55]	59.50	65.10	–	68.80	43.76	48.30	–	51.80
CTV-GS-I (ours)	58.88	65.07	61.57	68.36	43.11	49.60	44.52	53.21
CTV-VB-I (ours)	59.78	65.52	62.31	68.88	44.30	50.08	47.00	53.35
VLAD (PlacesNet)* [43]	66.27	66.12	54.70	67.61	51.50	48.97	40.58	51.73
CTV-GS-P (ours)	70.82	68.88	57.61	73.51	54.95	52.46	41.49	58.43
CTV-VB-P (ours)	70.90	68.73	58.13	73.88	55.16	52.95	43.34	58.39

* Our implementation.

Table III reports our results of proposed CTVs at different scale levels. At each single scale level, the proposed CTV-VB-I improves the classification accuracy up to 44.30% at the 256×256 scale level, 50.08% at the 128×128 scale level, and 47.00% at the 64×64 scale level. The performance of CTV-GS-I on each single scale level keeps pace with that of CTV-VB-I. The difference between them is only 0.48% at least and 2.48% at most. As mentioned above, this difference is reduced to 0.14% in the multi-scale case. But we cannot leave out one point that Variational Bayesian methods to solve CTM will take too much time when convergence, especially for a large-scale dataset, e.g., SUN 397. The GS based solution [47] facilitates the efficient CTV-GS-I. In general, our CTV-GS-I is computationally more efficient than CTV-VB-I, with classification performance being neck and neck with it.

The proposed CTV-VB-I has improvements of 4.73% at the 256×256 scale level, 4.74% at the 128×128 scale level, and 6.79% at the 64×64 scale level, in comparison with VLAD [43]. Similarly, CTV-GS-I outperforms VLAD by 4.04% on average. CTV-VB-I and CTV-GS-I work better than the Semantic FV at the 128×128 scale. At the 256×256 scale level, CTV-VB-I is better than Semantic FV and CTV-GS-I has a comparable performance with it.

3) PlacesNet Based Results: To fully validate the proposed CTVs, we conduct experiments based on the scene-centric PlacesNet [23] which is different from the object-centric ImageNet [25]. The performance of CTV-Ps is 5% higher than that of CTV-Is, as shown in Table II. Both CTV-VB-P and CTV-GS-P achieve 58.39% and 58.43%. Compared with the CNN-P, CTV-VB-P and CTV-GS-P achieve 4.07% and 4.11% gains, respectively.

In particular, CTV-VB-P and CTV-GS-P have the same tendency that the smaller the scale is, the lower the performance is, as shown in Table III. This could be rooted in the differences of learned semantics between PlacesNet and ImageNet. ImageNet is object-centric so that the learned CNN features focus on the high-level object-oriented semantics/objects. This is in accord with the fact that objects often appear at small scales in scenes. It is rational that the proposed CTV-Is on small scales are superior to those on large scales. However, PlacesNet is scene-centric and the learned CNN features tend to the global scene-oriented semantics. The smaller the scale is, the less scene-level information it contains. It is reasonable that the proposed CTV-Ps have higher performance on larger scales.

With the open-source code released by [43] and simple replacement from CNN-I to CNN-P, as shown in Table III, multi-scale CNN-P based VLAD is lower by about 6.7% than our multi-scale CTV-Ps. Comparing with the case of CNN-I, our CNN-P based CTVs outperform the VLAD with a significant margin. Especially, at three individual scale levels, the performance of CNN-P based VLAD is severely influenced by the differences of learned semantics between PlacesNet and ImageNet. It presents a larger decrement from the 256×256 scale to 64×64 scale.

To further analyze CTVs, we present experimental results of the CTV-Is on test images. The first row of Fig. 7 shows the recognition examples of *village* scene images where the *buildings*, *sky*, *trees* and *rocks* coexist. The proposed CTV leverages correlated latent topics learned from word co-occurrence to describe this semantic co-occurrence and to alleviate the word ambiguity problem. Besides, these images are true positives for CTV but false negatives for CNN. Take the first image in the first row for example. It is correctly recognized as *village* with the CTV-GS while wrongly recognized as *castle* with the CNN feature. Below it, we display three *castle* images to observe how much the *village* image is similar to them. *Buildings*, *sky*, and *trees* coexist in both the *village* and the *castle* scene. Obviously, capturing semantic correlations is also limited for feature encoding because inter-class similarity presents great challenges for scene image features including CNN and latent semantic representations. The proposed CTV with respect to the global latent parameter β essentially promotes how visual words effect each latent topic, which is beneficial to identify the differences among scene categories. The reason is that one theme or topic is subject to the particular property of one scene. Buildings marked in green rectangles in Fig. 7 vary greatly across different scenes, e.g., *castle*, *abbey*, *construction site*, *slum*, and *kasbah*. In Fig. 8, regions from two *village* images in the first and the last second columns of Fig. 7 are shown. It is clearer that significant differences exist between the labeled regions of two categories and it motivates the exploration of the CTV. In addition, four more examples from other categories are shown in Fig. 9. These four images from four categories are true positives for the CTV but are falsely recognized as other categories by the CNN features.

4) MIT Indoor 67 Dataset: Main results on the MIT Indoor 67 have been provided in Table IV. Similar to SUN 397, we compare the proposed CTVs with the most relevant methods that include baseline CNN features and other methods

Fig. 7. Recognition results of the *village* scene. In the first row, images are true positives for the CTV but false negatives for the CNN features. In the three rows below them, images from different categories are falsely recognized as *villages* with the CNN features.

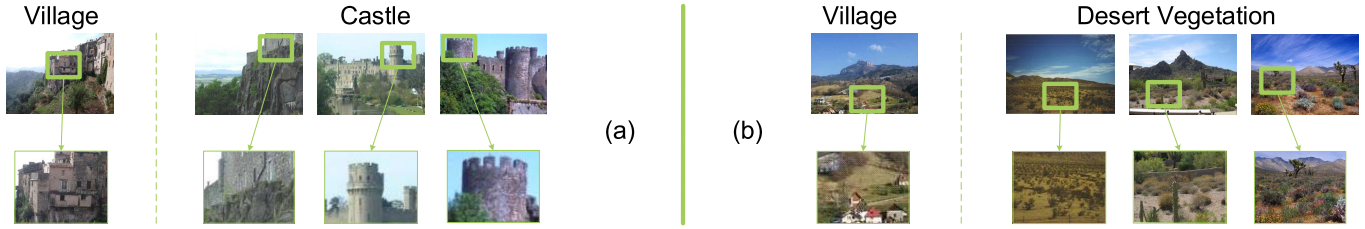


Fig. 8. Region examples.

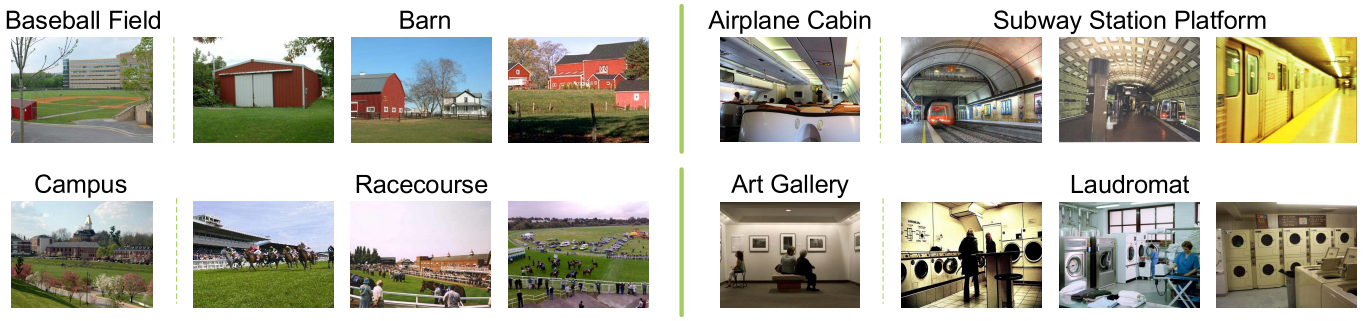


Fig. 9. Recognition results of four scene categories. For each category, we give one image which is the true positive of the CTV-GS but the false negative of the CNN features. Three images from the false negative categories of CNN are also shown.

derived from the Fisher Vector, and several other state-of-the-art methods. For these most relevant methods in Table IV, DMM FV [54] and Sparse Coding FV [56] are single scale, while VLAD [43], Semantic FV [54], CTV-GSs and CTV-VBs are multi-scale, i.e., three scales.

5) ImageNet Based Results: The first group of comparison methods involves CNN-I FC7 features as descriptors, except sparse coding FV [56] that utilizes CNN-I features of the sixth layer. Among these most relevant methods, a CNN baseline accuracy is 56.79% [23]. CTV-VB-I achieves 68.88% classification accuracy and CTV-GS-I obtains 68.36% performance. We come to the same conclusion on this dataset as SUN 397: CTV-VB-I and CTV-GS-I have the comparable performance and both of them significantly outperform the CNN-I baseline. The former is a bit more accurate while the latter has lower time complexity. What's more, the proposed CTV-Is are comparable to VLAD, DMM FV, Semantic FV, and Sparse Coding. Sparse Coding FV [56] extracts Fisher Vector of a sparse coding based model over local CNN-I features. Different from Semantic FV and VLAD, Latent GMM FV method [37] places a Dirichlet prior on mixing weights which are the parameters of GMM. It achieves 65.0% accuracy when the way of sampling patches is similar to ours, i.e., dense grid sampling. Due to Dirichlet prior, Latent GMM explicitly claims that each Gaussian component is independent of each other. In contrast, the proposed CTV-Is take correlations between

TABLE IV
COMPARISON ON THE MIT INDOOR 67 DATASET

Methods		Accuracy	Year	Description
Most Relevant Methods	CNN-I [23] (baseline)	56.79	2012	Deep networks trained on the ImageNet dataset
	CTV-GS-I (ours)	**68.36**	–	CTV based on CNN-I with Gibbs sampling, three scale levels
	CTV-VB-I (ours)	**68.88**	–	CTV based on CNN-I with Variational Bayesian, three scale levels
	CNN-P [23] (baseline)	68.24	2014	Deep networks trained on the Places dataset
	CTV-GS-P (ours)	**73.51**	–	CTV based on CNN-P with Gibbs sampling, three scale levels
	CTV-VB-P (ours)	**73.88**	–	CTV based on CNN-P with Variational Bayesian, three scale levels
	Latent GMM FV [37]	65.00	2015	Latent GMM based Fisher Vector; single scale
	Spare Coding FV [57]	68.20	2014	Sparse Coding based Fisher Vector; single-scale
	DMM FV [55]	68.50	2015	Dirichlet Mixture Model based Fisher Vector; single-scale
	Semantic FV [55]	68.80	2015	GMM Fisher Vector with natural parameterization; best three scale levels
	VLAD [43]	68.88	2014	VLAD concatenation of three scale levels
Other State-of-the-art Methods	Improved Object Bank [10]	46.60	2014	A large number of pre-trained object detectors
	DeCaF [26]	58.40	2014	Decaf; global CNN features
	FV + Bag of parts [58]	63.18	2013	GMM Fisher Vector; distinctive part detectors; part occurrences
	Mid-level elements [59]	64.03	2013	Mid-level visual element discovery as discriminative model seeking

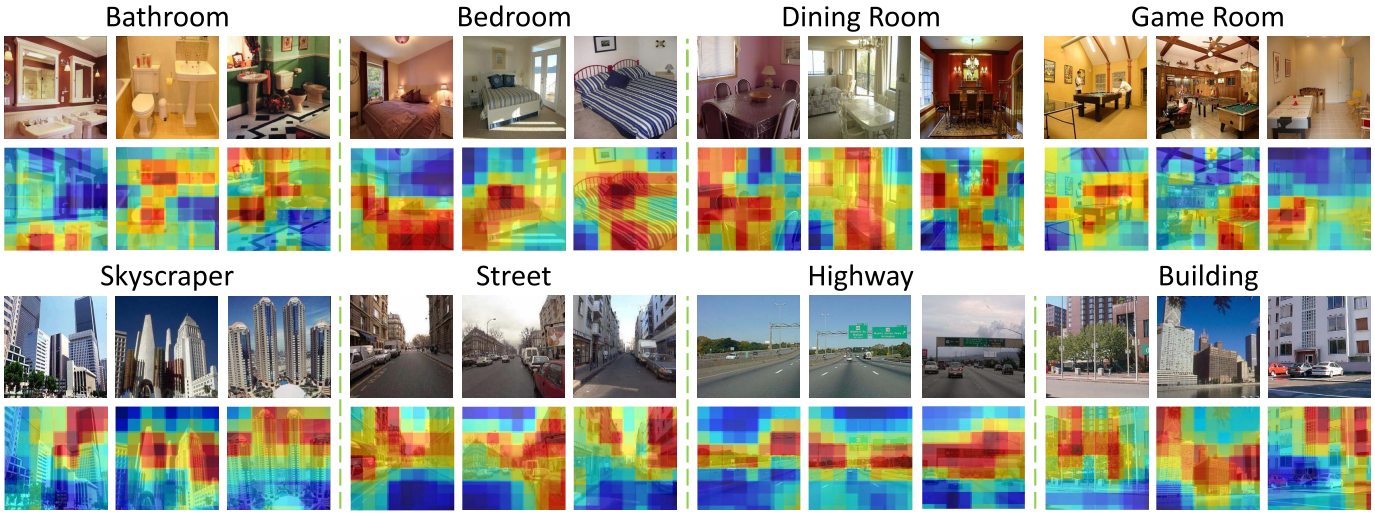


Fig. 10. Global topic probability map. It is interestingly observed that the topics of high-probability in the scene images usually correspond to correlated objects, e.g., *toilet* is correlated with *washbasin* in the *bathroom* scene, and *cushion* and *bed* in *bedroom* scene. (Best viewed in color.)

two components or clusters/themes/topics into consideration. CTV-VB-I outperforms it by 3.88%. Therefore, the independent assumption is too strict to characterize semantics for scene images.

We also evaluate the CTVs at different scale levels in Table III. At the 256×256 scale level, CTV-VB-I obtains 59.78% accuracy and significantly outperforms VLAD by 6.05%; CTV-VB-I achieves 58.88% which outperforms VLAD by 5.15%. At this scale level, the CTVs encode features from the whole image, rather than cropping the image into patches. The indoor scene images often present complex object configuration and within-class variation. Using feature encoding of cropped image patches could reduce the within-class variation but it is bound to severely destroy this object configuration when the patch size is too small. The reason could be that in small patches the descriptors tend to describe single objects or object parts rather the whole scene. This explains why CTV-Is extracted at a proper scale level, i.e., the 128×128 scale level, perform best.

6) PlacesNet Based Results: Based on the CNN-P (PlacesNet), the proposed CTVs achieve up to 73.88% accuracy as shown in Table III. Similar to the case of SUN 397, CTV-Ps at the individual scale prefer to the larger scale. CTV-VB-P achieves 70.90% classification accuracy at the 256×256 scale. This result of the 256×256 scale outperforms that of 128×128 scale with gains 2.17%, and that of 64×64 scale with gains 12.17%. This shows the similar trends in the SUN 397 dataset with the same essential reasons as discussed above.

C. Evaluation of Models and Parameters

1) Learned Topics and Correlation Between Topics: To further demonstrate the learned topics, we utilize a portion of SUN 397 images, which are fully annotated for object segmentations [30]. We name these images as SUN-anno dataset for simplicity and evaluate learned topics and topic correlations on this dataset. In detail, we utilize manually segmented object regions to build Topic Probability Maps (TPMs), which present the correspondence between topics and objects, and explicitly show what the learned topics are. The TPM is built by assigning each pixel of one image to a topic probability vector based on the learned global parameter β and

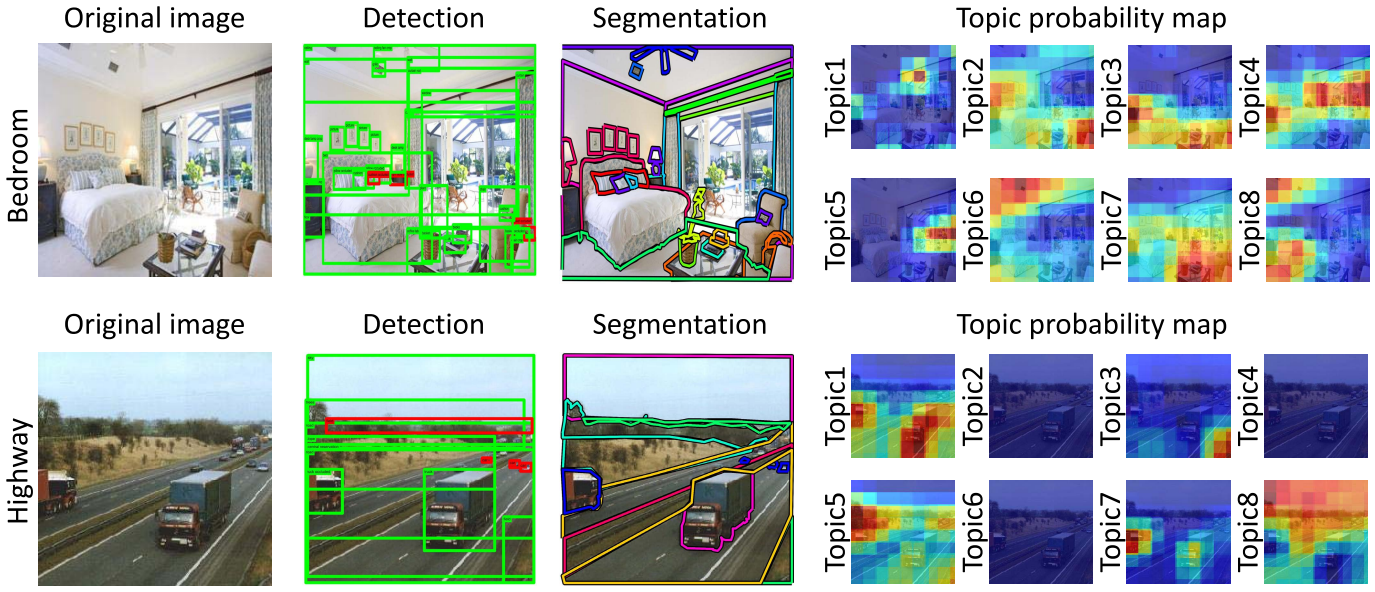


Fig. 11. Topic probability maps and their correspondence to detection and segmentation annotations. It is interestingly observed that the learned topics well correspond to the explicit objects and segmented image regions. In the *bedroom* scene image, high-probability regions of Topic2 well correspond to the *desk lamp* and *chair*. High-probability regions of Topic3 correspond to the *table* and *night table*. Topic4 is for *window*, Topic5 for *chair*, Topic6 for *ceiling*, Topic7 for *table*, and Topic8 for *ceiling*, *floor*, and *bed*. In the *highway* scene image, Topic1 is for a *truck* and an *occluded truck*, Topic3 for *field*, Topic5 for *tree*, Topic7 for *occluded truck*, and Topic8 for *sky*.(Best viewed in color.)

image-specific parameter ϕ, defined in Sec. III. Pixels in one patch share the same topic probability vector. As we are aware that a pixel may be shared for several patches, for a certain topic, we sum the corresponding probability values for each pixel across all the patches and obtain a TPM through normalization.

In Fig.10, we first show a global TPM which sums TPMs of topics in individual images from indoor categories and outdoor categories. It can be seen that the pixels of high topic probabilities in the TPMs well correspond to the segmented objects in the scene images. Moreover, the interesting objects have been spotted together, e.g., objects of *toilet* and *washbasin* emerge in the *bathroom* scene, objects of *cushion*/*window* and *bed* coexist in the *bedroom* scene, and objects of *building* and *car* coexist in the *street* scene.

As shown in Fig. 11, the TPMs of some topics exhibit an interesting correspondence between learned topics and annotated/detected objects. For example, in the *bedroom* scene, two high probability regions of Topic2 well correspond to the *desk lamp* and *chair*. Topic3 is for *table* and *night table*, Topic4 for *window*, Topic5 for *chair*, Topic6 for *ceiling*, Topic7 for *table*, and Topic8 for *floor* and *bed*. Unsurprisingly, these results demonstrate the capability of latent topic generation of CTV. Be aware that, even though the topics are shared among all the images, a topic would exhibit different object semantics in different images because of inferred image-specific parameter ϕ. The learned latent topics can discover the dominated explicit objects without any supervision of object segmentation or detection annotation, and present a category-specific property on the scene images. Thus, it promotes the proposed CTV features to be more discriminative and representative.

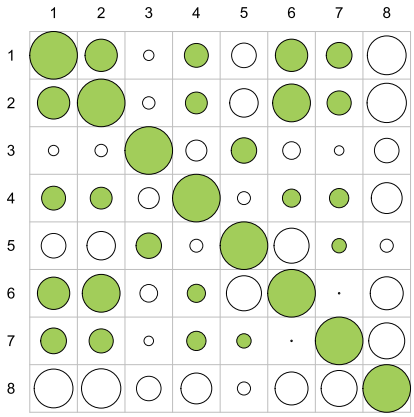


Fig. 12. Topic correlation matrix for 8 topics learned on the SUN-anno dataset. Solid circle stands for positive correlation between two topics, while open circle represents negative correlation between two topics. Larger radius, larger positive/negative correlations.

To explicitly show the topic correlations, the learned topic correlations matrix for 8 topics has been shown in Fig. 12, while the topic correlations with the overall statistical object results for all the learned topics have been shown in Fig. 13. As expected, the results can explain the positive/negative correlations to some extent. For example, object *chair* in Topic1 and object *table* in Topic2 often coexist, especially in the indoor scene images; this accords with the results that Topic1 has a positive correlation with Topic2 as shown in Fig. 12.

The coexistence of objects in many scenes and the subtle correspondence between objects and learned topics validate that correlations do exist in scene images. Therefore, the

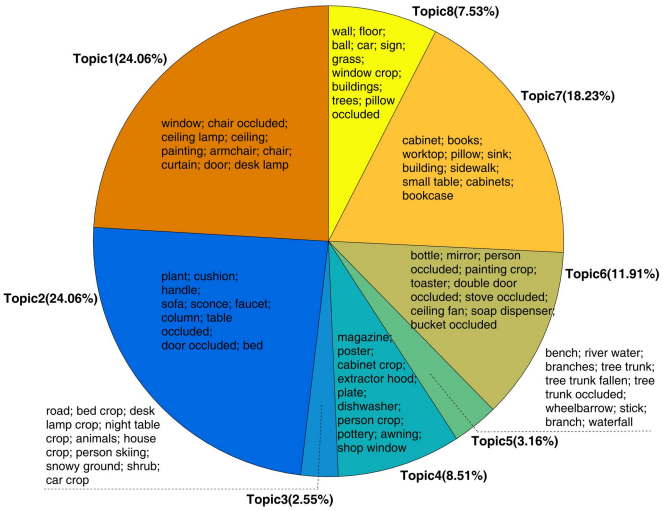


Fig. 13. The percentage of objects corresponding to one topic with a maximum probability. And the top ten object categories are also provided for each topic in the figure. It is observed that object *chair* in Topic1 and object *table* in Topic2 often coexist, especially in the indoor scene images. And this accords with the results that Topic1 has a positive correlation with Topic2 as shown in Fig.12. Best viewed in color and under zoom.)

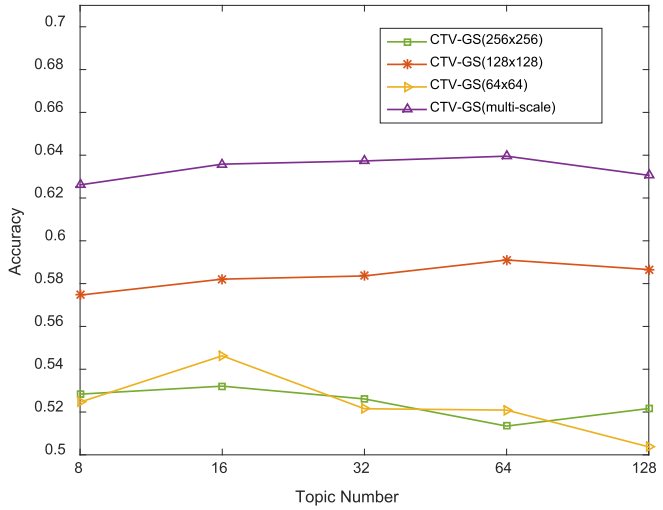


Fig. 14. Evaluation of topic numbers.

learned topics incorporate category-specific correlations and their rich semantics can enforce the within-class similarity. Together with the inter-class discriminative capability enforced with the Fisher Vector method, CTV demonstrates a great potential to the scene classification task.

2) Number of Topics: We evaluate the effect of topic numbers on the MIT Indoor 67 dataset. The results are present in Fig. 14. With the topic numbers ranging from 8 to 128, the performance of the multi-scale CTV-GS changes slightly, showing that the number of topics is not a main factor for CTV. The reason is that the CTVs are derived from latent correlated topics/semantics and that the CTVs with the limited topic number are capable of capturing the subtle differences of topics and words for each image in the Fisher Kernel space.

3) Solving Algorithms: We conduct experiments with the proposed features derived from two different CTM solving algorithms, CTV-VB and CTV-GS. In Table II, III and IV, it can be observed that, at either multi-scale or a single scale level, CTV-GSs are always neck and neck with CTV-VBs. In the case of three individual scales levels, the average difference between CTV-VB-I and CTV-GS-I is only 0.70% on the MIT Indoor 67 dataset and 1.38% on the SUN 397 dataset. At multi-scale levels, it decreases to 0.52% on the MIT Indoor 67 dataset and 0.06% on the SUN 397 dataset, respectively. The negligible performance difference shows that the Gibbs Sampling algorithm is a good approximation to the Variational Bayesian algorithm.

V. Conclusion

In the paper, we propose CTV representation for scene classification targeting to utilize the correlation among topics. By removing i.i.d assumption for local image patches and involving the logistic normal prior distribution, this method can better model the learning for features. Implemented on rich semantic information of CNN features, we explore underlying correlated semantics and encode them with the Fisher Vector framework to increase the discriminative capability. To make the method suitable for the large-scale datasets, we further provide a Variational Bayesian solution and a Gibbs sampling solution. The proposed CTV can be treated as an evolution oriented from the Fisher Vectors based on GMM or LDA. Experiments on large-scale datasets validate the effectiveness of CTV, showing its great improvement over CNN features and great potential to other Fisher Kernel based deep features. Together with GMM based Fisher Vector and LDA based Fisher Vector, our proposed CTV constructs a more complete generative model for image semantic representations.

Appendix

We provide the derivation details of CTV discussed in Section III-B.

Parameters of CTM are $\Theta = \{\mu, \Sigma, \beta\}$. The approximated log-likelihood of image d is L_{VB}:

$$\begin{aligned} L_{VB} &= E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N_d} E_q[\log p(z_n|\eta)] \\ &\quad + \sum_{n=1}^{N_d} E_q[\log p(w_{d,n}|z_n, \beta)] + H(q) \\ &= 1/2 \log|\Sigma^{-1}| - K/2 \log 2\pi \\ &\quad -1/2[Tr(diag(v_d^2)\Sigma^{-1}) + (\lambda_d - \mu)^T \Sigma^{-1}(\lambda_d - \mu)] \\ &\quad + \sum_{n=1}^{N} \{\sum_{i=1}^{K} \lambda_{d,i}\phi_{d,ni} - \zeta^{-1}(\sum_{i=1}^{K} \exp(\lambda_{d,i} + v_{d,i}^2/2)) \\ &\quad +1 - log\zeta\} + \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{d,ni} \log \beta_{i,w_{d,n}} \\ &\quad + \sum_{i=1}^{K} 1/2(\log 2\pi + \log v_{d,i}^2 + 1) \\ &\quad - \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{d,ni} \log\phi_{d,ni}, \end{aligned} \tag{10}$$

where ζ is a new introduced variational parameter. $\{\lambda, \nu, \phi\}$ are variational parameters as discussed in Section III and they are vectors with K elements which are indexed by i, $i = 1, ..., K$, for the image d.

The derived CTV $\varphi_{[\Theta]} = {I_{[\Theta]}}^{-1/2} u_{[\Theta]}$. The Fisher score $u_{[\Theta]} = \partial L_{VB}/\partial\Theta$ is the partial derivative of the likelihood with respect to parameters of CTM. Fisher information matrix is the second moment of the log-likelihood. Since the expectation of Fisher score is equivalent to zero, $I_{[\Theta]}$ is also the variance of Fisher score: $I_{[\Theta]}=E[{u_{[\Theta]}}^T u_{[\Theta]}]$. Under certain regularity conditions, the Fisher information is the negative of the expectation of the second derivative with respect to Θ: $I_{[\Theta]} = -E[\partial^2 L_{VB}/\partial\Theta^2]$. So we first compute Fisher score $u_{[\Theta]}$. The terms involving hyper-parameter μ in L_{VB} are:

$$L_{VB}^{[\mu]} = 1/2(\lambda_d - \mu)^T \Sigma^{-1}(\lambda_d - \mu). \tag{11}$$

The terms involving hyper-parameter Σ in L_{VB} are:

$$\begin{aligned} L_{VB}^{[\Sigma]} = {} & 1/2(\log|\Sigma^{-1}| + Tr(diag(\nu_d^2)) \\ & +(\lambda_d - \mu)^T \Sigma^{-1}(\lambda_d - \mu)). \end{aligned} \tag{12}$$

The terms involving global parameter β in L_{VB} are:

$$L_{VB}^{[\beta]} = \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{d,ni} \log \beta_{i,w_{d,n}}. \tag{13}$$

Now the Fisher scores of Equations (4)-(6) can be simply derived from Equations (11)-(13). Following $I_{[\Theta]} = -E[\partial^2 L/\partial\Theta^2]$, we then compute the second order derivative of Equations (11)-(13) for the Fisher information matrix. The results are Equations (7)-(9).

REFERENCES

[1] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.

[2] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2004, pp. 1401–1408.

[3] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947–963, Sep. 2001.

[4] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.

[5] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Content-based hierarchical classification of vacation images," in *Proc. IEEE Int. Conf. Multimedia Comput. Sys.*, Jun. 1999, pp. 518–523.

[6] C. Siagian and L. Itti, "Gist: A mobile robotics application of context-based vision in outdoor environment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Sep. 2005, p. 88.

[7] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Auton. Robots*, vol. 18, no. 1, pp. 81–102, Jan. 2005.

[8] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 883–890.

[9] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[10] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014.

[11] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 359–372.

[12] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.

[13] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 517–530.

[14] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[15] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 902–917, May 2012.

[16] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1999, pp. 50–57.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[19] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2006, pp. 147–154.

[20] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 17–35, 2007.

[21] J. Aitchison, "The statistical analysis of compositional data," *J. Roy. Statist. Soc. B (Methodological)*, vol. 44, no. 2, pp. 139–177, 1982.

[22] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1575–1589, Sep. 2007.

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[26] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn.*, Jun. 2014, pp. 647–655.

[27] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.

[28] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Eur. Conf. Comput. Vis. Workshop*, May 2004, pp. 1–2.

[29] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[30] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN database: Exploring a large collection of scene categories," *Int. J. Comput. Vis*, vol. 119, no. 1, pp. 3–22, Aug. 2016.

[31] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 487–493.

[32] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[33] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3304–3311.

[34] A. D. Holub, M. Welling, and P. Perona, "Combining generative models and Fisher kernels for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 136–143.

[35] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3278–3285.

[36] R. G. Cinbis, J. Verbeek, and C. Schmid, "Image categorization using Fisher kernels of non-iid image models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2184–2191.

[37] R. G. Cinbis, J. Verbeek, and C. Schmid, "Approximate Fisher kernels of non-iid image models for image categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1084–1098, Jun. 2016.

[38] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[39] J. Atchison and S. M. Shen, "Logistic-normal distributions: Some properties and uses," *Biometrika*, vol. 67, no. 2, pp. 261–272, Jan. 1980.

[40] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[41] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," Dept. Comput. Sci., Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep., 2001, vol. 1, p. 40.

[42] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974.

[43] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 392–407.

[44] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.

[45] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[46] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis*, vol. 105, no. 3, pp. 222–245, Dec. 2013.

[47] J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang, "Scalable inference for logistic-normal topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2445–2453.

[48] D. Mimno, H. M. Wallach, and A. McCallum, "Gibbs sampling for logistic normal topic models with graph-based priors," in *Proc. Adv. Neural Inf. Process. Syst. Workshops Analyzing Graphs*, 2008.

[49] T. Salimans, D. P. Kingma, and M. Welling, "Markov chain Monte Carlo and variational inference: Bridging the gap," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. 2015, pp. 1218–1226.

[50] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.

[51] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf.*, Nov. 2011, pp. 1–12.

[52] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3485–3492.

[53] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.

[54] M. Dixit *et al.*, "Scene classification with semantic Fisher vectors," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2974–2983.

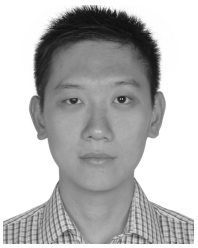
[55] A. Bergamo and L. Torresani, "Classemes and other classifier-based features for efficient object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1988–2001, Oct. 2014.

[56] L. Liu, C. Shen, L. Wang, A. van den Hengel, and C. Wang, "Encoding high dimensional local features by sparse coding based Fisher vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1143–1151.

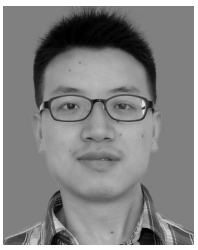
[57] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 923–930.

[58] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 494–502.

Pengxu Wei received the B.S. degree in computer science and technology from the China University of Mining and Technology, Beijing, China, in 2007. She is currently pursuing the Ph.D. degree with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences. Her current research interests include computer vision and machine learning, specifically for data-driven vision and scene image recognition.



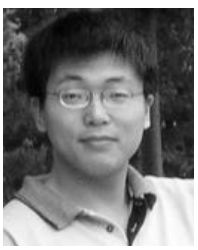
Fei Qin (M'05) received the B.S. degree in information engineering from the Huazhong University of Science and Technology, in 2004, the M.S. degree in electronic engineering from the Beijing Institute of Technology, in 2006, and the Ph.D. degree in electronic engineering from the University College London, U.K., in 2012. From 2006 to 2008, he served as a Product Manager with Crossbow Technology, Beijing Representative Office. He has been an Associate Professor with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, since 2015, where he is currently a Lecturer from 2012 to 2015. His current research interest is the joint optimization method of wireless network and information system.



Fang Wan received the B.S. degree in measurement and control technology and instrument from Wuhan University in 2013, and the M.S. degree in electronic and communication engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree in computer application technology. His research is on computer vision and machine learning, specifically for weakly supervised learning and object detection task.

Yi Zhu received the B.S. degree in software engineering from Sun Yat-sen University, Guangzhou, China, in 2013. She is currently pursuing the M.S. degree in computer science with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China. Her current research interests include object detection and image classification.

Jianbin Jiao (M'10) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively, all in mechanical and electronic engineering. From 1997 to 2005, he was an Associate Professor with HIT. Since 2006, he has been a Professor with the School of Electronic, Electrical, and Communication Engineering, University of the Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, pattern recognition, and intelligent surveillance.



Qixiang Ye (M'10–SM'15) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He has been a Professor with the University of Chinese Academy of Sciences since 2009, and was a Visiting Assistant Professor with the Institute of Advanced Computer Studies, University of Maryland, College Park, in 2013. He has authored over 50 papers in refereed conferences and journals, and received the Sony Outstanding Paper Award. His current research interests include image processing, visual object detection and machine learning. He pioneered the Kernel SVM-based pyrolysis output prediction software which was put into practical application by SINOPEC in 2012. He developed two kinds of piecewise linear SVM methods which were successfully applied into visual object detection.