Correlated Topic Vector for Scene Classification

Pengxu Wei, Fei Qin, 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 datasets. This paper proposes a novel generative image representation, Correlated Topic Vector (CTV), to model such semantic correlations. Oriented from correlated topic model, CTV intends to naturally utilize the correlations among topics which are seldom considered into the conventional feature encoding, e.g. Fisher Vector, but do exist in scene images. It is expected that the involvement of correlations may increase the discriminative capability of learned generative models and consequently improve the recognition accuracy. Incorporated with Fisher Kernel method, CTV inherits the advantages of Fisher Vector. The contributions to topics/themes of visual words have been further employed by incorporating the Fisher Kernel framework to indicate the differences among scenes. Combined with the deep CNN features and Gibbs sampling solution, CTV shows great potential when processing large-scale and complex image datasets. Experiments on two scene image datasets demonstrate that CTV improves significantly the deep CNN features, and outperforms existing Fisher Kernel based features.**

***Index Terms*—Generative feature learning, Correlated Topic Vector, semantic correlation, Fisher Kernel.**

I. Introduction

SCENE classification has been widely explored, promoting related computer vision research topics including object recognition [1], [2], image retrieval [3]–[5], and intelligent robot navigation [6], [7]. A scene image is composed of several semantic entities (e.g. *sky*, *rock*, *street*, *car*). These entities are often organized in unpredictable layouts [8], [9] and could be shared with multiple categories, which invite intra-class variability and extra-class similarity for scene recognition. Scene labels (e.g. *coast*, *village*, *coast*, *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 these semantics. The first class of semantics consists of object-centric approaches that model pre-defined explicit semantics, e.g. objects (*sky*, *water*, *grass* and so on) or scene categories (*village*, *forest*, *kitchen* and so on). 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 its holistic description with the aid of implicit themes. And it works without explicit image segmentation, manual theme annotation or extensive object detections.

The scene-centric representation is conventionally built on Bag-of-Words (BoW) which 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). This problem is demonstrated in Fig. 1. As shown in its first row, 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 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 which 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 appear together in the *village* scene. Obviously, a scene exhibits a strong semantic/theme correlation property, and more importantly this property is specific for a scene category distinguishing itself from others. Unfortunately, such correlation is ignored in most existing work besides BoW. For example, LDA imposes Dirichlet distribution prior on the topic proportions [18], which poorly assumes that themes/topics are independent from each other.

In this paper, we propose a new feature representation, named Correlated Topic Vector (CTV), that 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

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).

distributions [13], [14], fails to perform well consistently. This similar case happens to other topic models (e.g. pLSA and LDA) [22]. Besides, alone with the significantly increasing categories of scene images, topic models including CTM tend to fail to describe the intra-class variability and extra-class similarity. This problem is aggravated when using latent topic distributions as their unsupervised learning obscures differences among scene categories.

The proposed CTV further explores the contributions of low level visual words to the generating processes 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 word 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 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 efforts. It is an intuition that regarding CNN feature of the hidden fully connected layer as a learnt soft-assignment word histogram for a whole image as well as avoiding to build a vocabulary relying on CNN as local descriptors.

To summarize, this work has the following contributions:

1. We propose a new image representation, Corrected Topic Vector, which can capture the correlations among semantic topics of images.

2. We derive the formal expression of CTV in Fisher Kernel space to enhance discriminative capacity.

3. We implement CTV with CNN features and efficient Gibbs sampling solution to large-scale datasets.

In the remainder of the paper, we review related work in Section II and discuss the detail 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

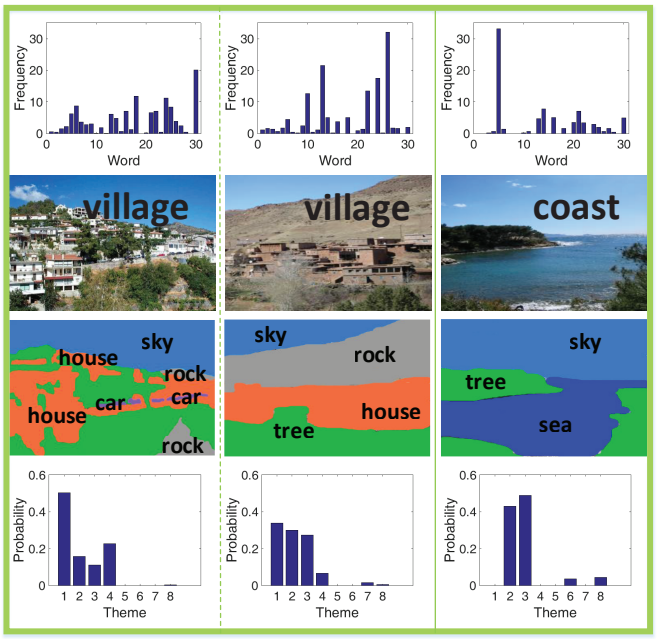

Fig. 1. Examples of *village* and *coast* scene images. In the first row, histograms of visual words are shown; Themes are provided for three images, among which two images of scene *village* and one image of scene *coast* present in the second row; Their corresponding theme probability distributions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

detection results of diverse objects. L. 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. It is hard to generalize OB to large-scale scene image sets such as SUN 397 [30] or Places 205 [23]. Besides, L. 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 denoted to Fisher Kernel [31] to improve BoW. Tommi S. Jaakkola and David Haussler [21] provide a formulation of Fisher Kernel for classification tasks. Florent Perronnin and Christopher 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]. VLAD (Vector of Locally Aggregated Descriptors) [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 Dirichlet distribution.

CNN features 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

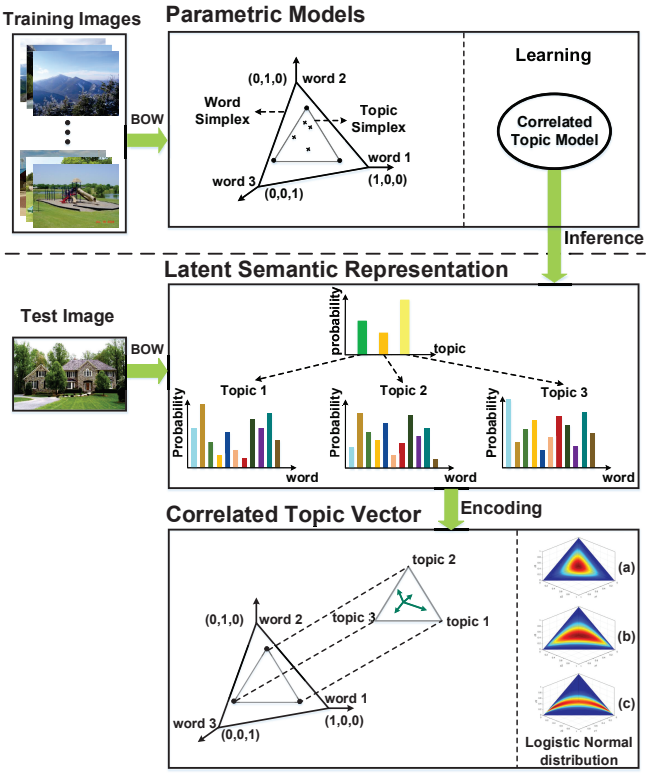

Fig. 2. Correlated Topic Model learning and Correlated Topic Vector encoding.

extracted from local patches and build VLAD embedding for image recognition. M. 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. With the help of Dirichlet Mixture Models (DMM), the DMM FV is induced as a more natural embedding than GMM FV. When local semantic descriptors are modeled as a multinomial distribution, the DMM FV is induced as a more natural embedding than GMM FV. Besides, the natural parameterization transformation alleviates highly non-Euclidean property of SMN descriptors. A semantic FV is then computed as a Gaussian Mixture FV in the space of the natural parameters.

A considerable number of works built on Fisher Kernel framework for image recognition have made great strides, however 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 characterics, and thus cannot always hold. In addition, semantic correlation is seldom considered in existing works. Considering an image as an unordered set of regions, R. G. 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 model 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

TABLE I
THE GENERATING PROCESS OF IMAGE d.

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) Firstly 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) Then for each topic, draw word $w_{d,n} \vert \{z_n, \beta\}$ from $Mult(\beta_{z_n})$.

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 of the wide exploration of latent semantic in these works, 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 framework may further enhance the discriminative capacity, the task of scene classification will be pretty straightforward: first 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 feed to one classifier, e.g. SVM in our implementation, 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 Bayesion (VB) method and Gibbs sampling (GS) method. The basic scheme of CTV encoding has been shown in Fig. 2.

A. Latent Semantic Representation

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

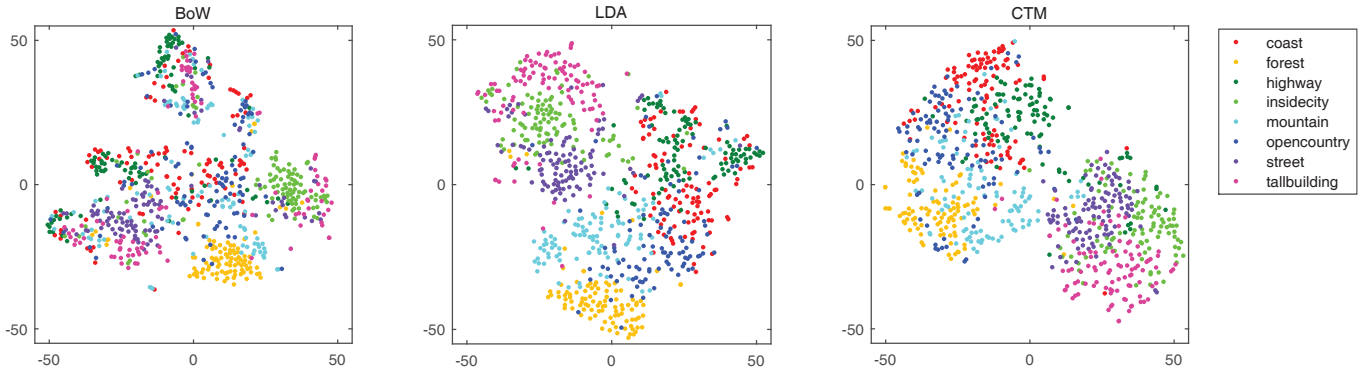

Fig. 3. Feature visualization using t-sne [38]. We visualize three types features on the SCENE 8 dataset, bow, latent semantic representations of LDA and CTM. It can observed that latent semantic features derived from LDA are more compact than BoW, while latent semantic features of CTM demonstrate a superior cluster effect for all the scene categories in semantic feature space. (Best viewed in color.)

is hinted from 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 modelled 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 $\theta_d = [\theta_d^1, ..., \theta_d^K]$, where

$$\theta_d^i = f(\eta_d^i) = \exp \eta_d^i / \sum\nolimits_j \exp \eta_d^j. \tag{1}$$

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) θ remain the image-specific property; (2) topic proportions θ imply correlation-ship among topics stemming from a logistic normal prior. As validated in Fig. 3, the latent semantic representation derived from CTM is more compact and discriminative than conventional BoW and LDA.

The performance of CTM based latent semantic representation can not consistently increase with the increasing number of topics yet. The experimental results shown in Fig. 4 validate this situation. Given 200 and 256 vocabulary sizes, the topic representations perform better and better, but its accuracy decreases when the topic number is larger than 50 and 60 respectively. A similar situation of LDA has been demonstrated in [22]. This limitation of latent semantic representations may be not resulted from poor statistical estimation, but from the intrinsic ambiguity of the underlying BoW representation [15]. Another reason is that latent semantic features stemmed from word co-occurrence fail to utilize the statistical

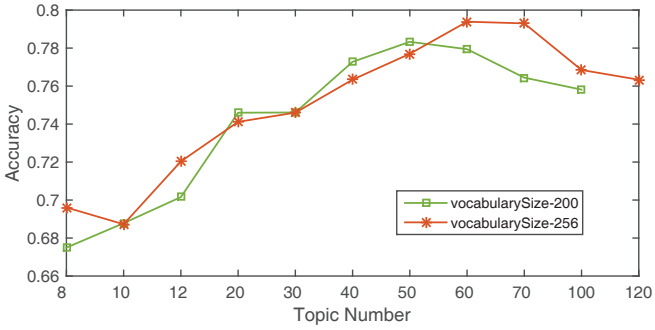

Fig. 4. Performance of CTM based latent semantic representation on the SCENE 8 dataset [40].

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 modelling of correlations among topics.

To further encode informative features based on semantics, we attempt to explore the contributions of low level words to the generating of middle level semantics from the information geometry view. Therefore, we propose the feature encoding scheme of Correlated Topic Vector (CTV) 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 by utilizing the Variational Bayesian method, which computes the gradient of the log likelihood of an image and the Fisher information matrix with respect to model parameters, i.e. hyper-parameters $\{\mu, \Sigma\}$ and global latent parameter β.

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

$$\begin{aligned} &p(w_d|\mu, \Sigma, \beta) \\ &= \int p(\eta|\mu, \Sigma)(\prod_{n=1}^{N_d} \sum_{z_n} p(z_n|\eta) p(w_{d,n}|z_n, \beta))d\eta, \end{aligned} \tag{2}$$

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 $p(z_n|\eta)$. As a result, it is hard to analytically compute the integrals in Equation (2). In other word, we cannot directly derive the gradient of the log likelihood to obtain CTV features.

We resort to the Variational Bayesian method [41] to derive the formal expression. Variational Bayesian is an approximate approach that optimizes a deterministic objective lower bounded on the data of log likelihood [41]. 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}:

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

where E is the expectation with respect to the variational distribution q whose parameters $\{\lambda,\nu^2,\phi\}$ are variated from $\{\mu,\Sigma\}$, and $H(q)$ denotes the entropy of this distribution. Details on how to obtain $\{\lambda,\nu^2,\phi\}$ from $\{\mu,\Sigma\}$ can be found in [20].

Now, we could derive the formal expression of the correlated topic vector for image d based on the learned parameters of CTM model, i.e. Θ= $\{\mu,\Sigma,\beta\}$. Its form is $\varphi_{[\Theta]} = {I_{[\Theta]}}^{-1/2} u_{[\Theta]}$. $u_{[\Theta]} = \partial L/\partial\Theta$ denotes as Fisher score that 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 $\{\mu,\Sigma\}$ and global parameter β are

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

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

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

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

μ and Σ are parameters of the true multivariate Gaussian distribution, and they should be learnt 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. ϕ_{ni} is a multinomial parameter and denotes how likely a word $w_{d,n}$ occurs given topic assignment z. $u_{[\beta]}$ can be regarded as the expectation of word occurrence whose possibility ϕ_{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 Appendix A.

Fisher information matrix can be expressed as:

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

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

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

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 these three vectors and feature normalization (e.g. power normalization and L_2-normalization [42] [43]).

C. Gibbs Sampling Solution

Variational Bayesian methods mentioned above maximizes an explicit objective, and can be utilized to derive the formal expression of the CTV. We denote CTV derived from the VB approach as CTV-VB. However, as demonstrated in [44] and [45], the costly iterative optimization computation makes it impractical to scale it to large and complex datasets. To address this limitation, we further resort to the scalable Gibbs sampling algorithm [44] and name the resulted CTV as CTV-GS. More importantly, we aim to explore the CTV derivation with Gibbs sampling solution.

Gibbs sampling avoids explicit 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} [46], $L_{GS} = L_{VB} - E_{q(z_T|w)}\{D_{KL}[q(y|z_T,x)||r(y|z_T,x)]\} \leq L_{VB}$, 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 [46]. 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, parameters involved in CTV are learnt with Gibbs sampling and the encoding of CTV features relys on the Variational Bayesian method. Both Variational Bayesian and Gibbs sampling methods are approximations of the log probability of CTM, which characterize the same dependence among variables in a hierarchical graphic model. The feasibility of such approximation method is demonstrated by the experimental results in Section IV.

D. CNN-based Implementation

Deep CNN has demonstrated remarkable recognition performance [24], [26], [47]. Especially, its activated features of later layers present excellent generalization of image representation and powerful semantic clustering results [26].

We evaluate the proposed CTV based on local descriptors extracted using CNN [24]. CNN features of later layers are considered as a type of soft-assignment BoW. We build our deep-BoW based on them. In detail, we divide one image into regions sampled on a dense grid with $P \times P$ pixels and S-pixel stride size. Fully connected CNN outputs of the seventh layer, denoted as FC7, are extracted for each region. we average these local descriptors across regions. Followed by numerical truncation, the feature pooling, e.g. average pooling and max pooling, aids to obtain deep-BoW features with desired dimensions. One popular method named as CNN-BoW in this paper attempts to encode BoW by replacing local descriptors e.g. SIFT with CNN [11], [37], [48]. Compared with it, our deep-BoW can avoid a series of costly clustering and quantization operations. Besides, it accords with the evidence that a scene consists of objects. CNN is trained on an object-centric dataset with 200 categories, ImageNet [25]. After multi-hierarchy feature encoding and pooling, the activated features of fully connected layers with rich object information, can be regarded as learned soft-assignment histogram of visual words without extra object detections. Besides, as pointed in [48], the rectified linear unit (ReLU) transformation guarantees all the values of FC7 are non-positive for encoding deep-BoW. Since some objects often appear in different regions in a scene, features extracted from each region followed by pooling may be more representative than those from the whole image.

In Fig. 5, we provide experimental comparison results for CNN-BoW and deep-BoW with different dimensions on the MIT Indoor 67 dataset [49]. CNN-BoW derived from clustering methods e.g. GMM and k-means, performs worse than deep-BoW. We consider three different pooling methods to obtain lower dimension deep-BoW: fixed pooling, random pooling and max pooling. With fixed pooling, we select elements of the FC7 feature vector at different dimensions every fixed step stride. With random pooling, we select elements randomly. With max pooling, every certain step strides, we choose the dimension with maximum value among elements in each step stride. No matter which pooling approach it deploys, deep-BoW is always better than CNN-BoW, with the same dimensions which range from 256 to 4096. One crucial reason is that dropout for CNN ensures that three pooling above perform stable.

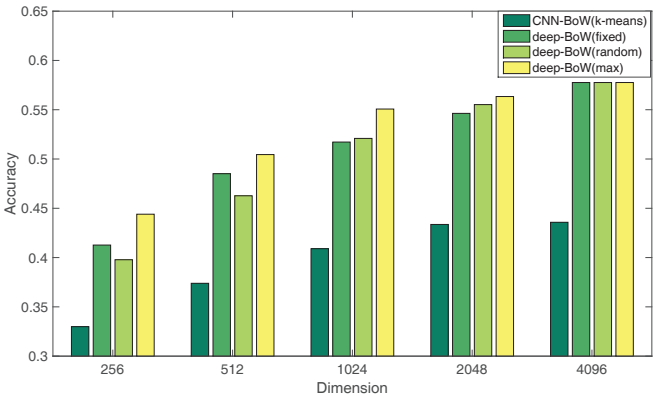

Fig. 5. Comparison of CNN-BoW and deep-BoW on the MIT Indoor 67 dataset.

IV. Experiments

In this section we report experimental results designed to evaluate the performance of the proposed CTV. Similar to [32], [43], [50], we train one-vs-all linear SVM classifiers on training images with LIBSVM [51]. The evaluation metric is average classification accuracy [48].

A. Setup

Datasets. We conduct experiments on two benchmark datasets: MIT Indoor 67 [49] and SUN 397 [30], [52]. MIT Indoor 67 contains 67 scene categories. Images have been split into 5360 training images and 1340 testing images, i.e. 80 training and 20 test images per category. 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.

Proposed Methods. We implement CTV-VB and CTV-GS (denoted as CTVs for convenience), on three scale levels mentioned above respectively. 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. Since Fisher information matrix is immaterial as pointed in [22], we approximate CTV with its Fisher score.

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. Patches are sampled with the stride of 32 pixels on all the scale levels. CNN FC7 features are extracted using the Caffe package [47] pre-trained on the ImageNet dataset [25]. To learn involved parameters for CTV, the deep-BoW is fed to the VB based CTM or GS based scalable CTM.

B. Main Results

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 (1) most relevant methods derived from the Fisher Vector framework, and with (2) other state-of-the-art methods.

TABLE II
COMPARISON ON THE SUN 397 DATASET.

	Methods	Accuracy	Year	Description
Most Relevant Methods	CNN [23]	42.61	2012	Networks trained on ImageNet image dataset
	DMM FV [53]	49.86	2015	Dirichlet Mixture Model based Fisher Vector
	Semantic FV [53]	51.80	2015	GMM Fisher Vector, natural parameterization, best three scales
	VLAD [48]	51.98	2014	the concatenation of VLAD on three scale levels
	CTV-GS(ours)	**53.21**	–	Gibbs sampling based Correlated Topic Vector
	CTV-VB(ours)	**53.35**	–	Variational Bayesian based Correlated Topic Vector
Other State-of-the-art Methods	SPMSM [11]	28.20	2012	Spatial pyramid matching, predefined semantic themes
	Meta-classes [54]	36.80	2014	Classifier-based features
	SUN(MKL) [52]	38.00	2010	Multi-kernel learning
	DeCaF [26]	40.94	2014	DeCaf, global features
	Places-CNN [23]	**54.32**	2014	Networks trained on a large-scale scene image dataset

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 [48]	53.73	65.52	62.24	68.88	39.57	45.34	40.21	51.98
Semantic FV [53]	59.50	65.10	–	68.80	43.76	48.30	–	51.80
CTV-GS(ours)	58.88	65.07	61.57	68.36	43.11	**49.60**	**44.52**	**53.21**
CTV-VB(ours)	**59.78**	**65.52**	**62.31**	**68.88**	**44.30**	**50.08**	**47.00**	**53.35**

TABLE IV
COMPARISON ON THE MIT INDOOR 67 DATASET

	Methods	Accuracy	Year	Description
Most Relevant Methods	CNN [23]	56.79	2012	Networks trained on ImageNet image dataset
	Latent GMM FV [37]	65.00	2015	Fisher Vector based on Latent GMM; grid sampling patches
	Spare Coding FV [55]	68.20	2014	Sparse Coding based Fisher Vector
	DMM FV [53]	68.50	2015	Dirichlet Mixture Model based Fisher Vector
	Semantic FV [53]	68.80	2015	GMM Fisher Vector, natural parameterization, best three scales
	VLAD [48]	68.88	2014	VLAD Concatenation of three scale levels
	CTV-GS(ours)	68.36	–	Correlated Topic Vector with Gibbs sampling solution
	CTV-VB(ours)	**68.88**	–	Correlated Topic Vector with Variational Bayesian solution
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 features
	FV + Bag of parts [56]	63.18	2013	GMM FV; distinctive part detectors; part occurrences
	Mid-level elements [57]	64.03	2013	Mid-level visual element discovery as discriminative model seeking
	Places-CNN [23]	68.24	2014	Networks trained on Scene image dataset

The first group of comparison methods involves CNN FC7 features as descriptors. These most relevant methods include baseline CNN features and several Fisher Vector based methods: DMM FV [53], Semantic FV [53] and VLAD [48]. DMM FV relys on DMM to model semantic multinomial descriptors. Semantic FV improves DMM FV with natural parameterizations of multinomial parameter vectors and computes Fisher Vectors with learnt GMM. VLAD is pointed as an approximation of Fisher Vector based on GMM [53]. Among these most relevant methods, a CNN baseline accuracy of 42.61% has been achieved as reported in [23]. In the table II, the results of VLAD [48], Semantic FV [53], CTV-GS and CTV-VB are multi-scale. CTV-VB achieves 53.35% and CTV-GS achieves 53.21%. Both of them achieve a gain of up to 10.74% and 10.60%, compared with CNN. The difference of 0.14% between them demonstrates the validness of approximation of CTV-VB by CTV-GS. This issue will be discussed again when evaluating features at different scales. DMM FV [53] is built on DMM which assumes that themes/topics are independent from each other. Its performance is less by 3.49% than CTV-VB and by 3.35% than CTV-GS. Besides, CTV-VB and CTV-GS achieve respectively 1.55% and 1.41% gains than Semantic FV. The concatenation of Semantic FVs at best four scales achieves 53.0% [53], which is lower than CTV-VB and CTV-GS. Gong *et al.* [48] report that the CNN based multi-scale VLAD can improve the classification performance up to 51.98%. CTV-GS and CTV-VB also outperform it. This result validates the i.i.d. assumption may not enough to model topics as it is pointed that VLAD is regarded as the approximation of Fisher Vector based on GMM [53].

Table III reports our results of proposed CTVs at different scale levels. For each single scale level, the proposed CTV-VB improves the classification accuracy up to 44.30% for 256×256 scale level, 50.08% for 128×128 scale level, and 47.00% for 64×64 scale level. The performance of CTV-GS on each single scale level keeps pace with that of CTV-VB. The difference between them is only 0.48% at least and 2.48% at most. As mentioned before, 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. It will be impractical

Fig. 6. Recognition results of the *village* scene. In the first row, they are true positive for the CTV but false negative for the CNN features. In the three rows below them, these three images are from the scene category that they are wrongly recognized as with CNN features respectively.

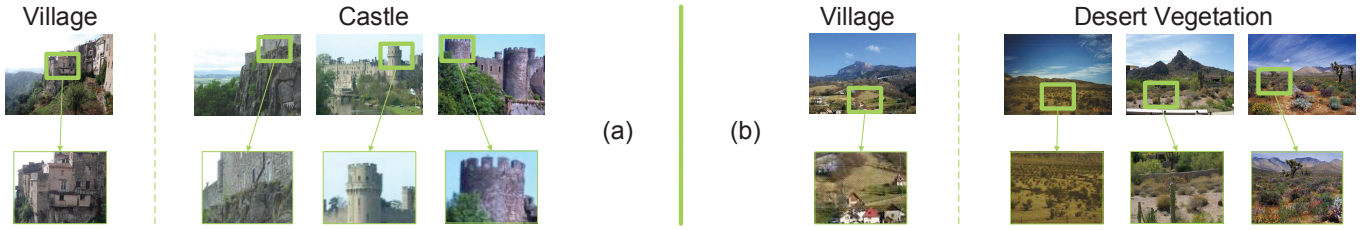

Fig. 7. Region examples.

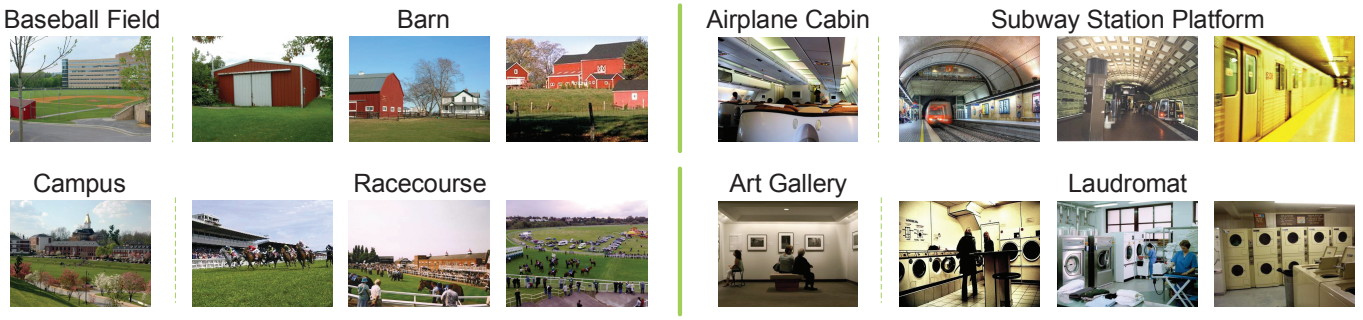

Fig. 8. Recognition results of four scene categories. For each category, we give one image which is the true positive of the CTV-GS but the false negative of CNN feature, and three negative images from a very similar category.

for implementations as argued in [44]. In general, our CTV-GS is computationally more efficient than CTV-VB, with classification performance being neck and neck with it.

The proposed CTV-VB has improvements of 4.73% for the 256×256 scale level, 4.74% for 128×128 scale level, and 6.79% for the 64×64 scale level, in comparison with VLAD [48]. Similarly, our CTV-GS also outperforms VLAD by 4.04% on average. As for Semantic FV, both CTV-VB and CTV-GS work better on 128×128 scale than Semantic FV. For 256×256 scale, CTV-VB is better than Semantic FV and CTV-GS has a comparable performance with it.

To further analysis CTVs, we present experimental results on test images. Fig. 6 shows the recognition examples of *village* scene images. Among *village* scene images shown in the first row, *buildings*, *sky*, *trees* and *rocks* almost appear together. The proposed CTV leverages correlated latent topics learnt from word co-occurrence to describe this semantic co-occurrence and to eliminate 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* scene category with the CTV-GS while it is wrongly recognized as *castle* scene category with the CNN feature. Below it, we also display three representative images whose category is *castle* to observe

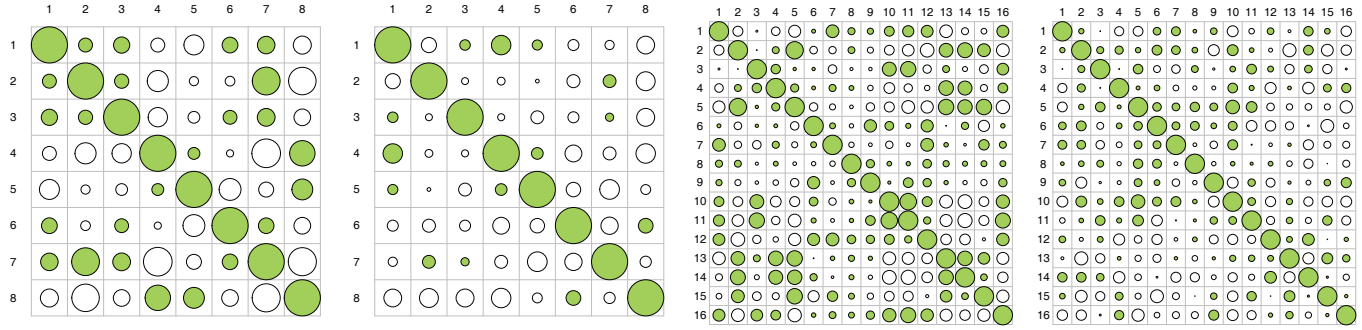

Fig. 9. Topic correlation matrix. The first and second columns are those for 8 topics on the SUN 397 dataset and MIT Indoor 67 dataset, and the last two are for 16 topics on both datasets. Solid circle stands for positive correlation between two topics, while open circle represents negative correlation between two topics. Larger radius, larger positive/negative correlation.

how much this *village* image is similar to them. *Buildings*, *sky*, and *trees* all appear together in the *village* and *castle* scene. Obviously, capturing semantic correlation is also limited for feature encoding and extra-class similarity brings 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. 6 vary greatly across different scenes, e.g. *castle*, *abbey*, *construction site*, *slum*, *kasbah*. In Fig. 7, regions from two *village* images in the first and the last second columns of Fig. 6 are shown. It is more clear that large differences exist between the labelled regions of two categories and it motivates the exploration of the CTV. In addition, four more examples from other categories are shown in Fig. 8. These four images from four categories are true positives for the CTV but are incorrectly recognized as other categories whose three representative images are present in Fig. 8.

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 (1) most relevant methods that include baseline CNN features and other methods derived from the Fisher Vector framework, and with (2) other state-of-the-art methods.

The first group of comparison methods involves CNN FC7 features as descriptors, except Sparse Coding FV [55] that utilizes CNN features of the sixth layer. Among these most relevant methods, a CNN baseline accuracy is 57.69%. CTV-VB achieves 68.88% classification accuracy and CTV-GS obtains 68.36% performance. We come to the same conclusion on this dataset as SUN 397: CTV-VB and CTV-GS have the comparable performance. The former is a bit more accurate while the latter costs less time complexity. What's more, our CTVs are comparable to VLAD, DMM FV, Semantic FV and Sparse Coding. Sparse Coding FV [55] extracts Fisher Vector of a sparse coding based model over local CNN featuers. 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 to each other. Contrary to it, our CTVs take correlations between two components or clusters/themes/topics into consideration. CTV-VB outperform it by 3.88%. Therefore, the independent assumption is strict to character semantics for scene images.

We also evaluate the CTVs at different scale levels in Table III. For the 256×256 scale level, CTV-VB obtains 59.78% accuracy and outperforms VLAD by 6.05%; CTV-VB also achieve 58.88% and is better than VLAD by 5.15%. On the 256×256 scale level, it encodes features just from the whole image, rather than cropping the image into patches. The indoor scene images often present more complex objects configuration because humans interact strongly in the places where they stay. Cropped patches may be robust to the intra-class variability (e.g. different sight ranges) but it is bound to severely destroy this configuration when keeping decreasing the patch size since the descriptors tend to describe one object or parts rather than the whole scene. This explains why CTVs extracted on 128×128 scale level perform best among three scales and why the performances of CTVs decrease on the 64×64 scale. In general, our CTVs still performs well on MIT Indoor 67 and are comparable with VLAD, Semantic FV.

C. Evaluation of Parameters

Correlation between topics: We visualize correlation matrixes obtained on the MIT Indoor 67 dateset in Fig. 9. With the aid of CTM, the underlying correlated structure among topics is captured.

Number of topics: We evaluate the effect of topic numbers on the MIT Indoor 67 dataset. The results are present in Fig. 10. With the topic numbers ranging from 8 to 128, the performance of the multi-scale CTV-GS change slightly. The similar cases occur on the 256×256, 128×128 and 64×64 scale levels. We can see that the number of topics may be not a main factor for the encoding of CTV.

Solving algorithms: we conduct experiments with the proposed features derived from two different CTM solving algorithms, CTV-VB and CTV-GS. From Table II, III and IV,

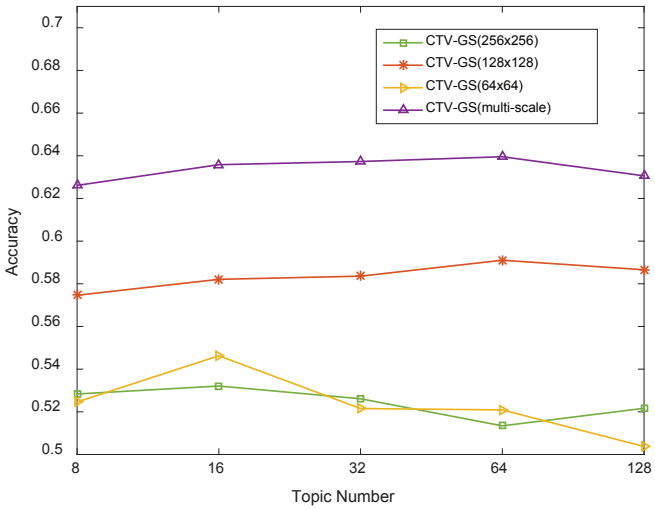

Fig. 10. Evaluation of topic numbers on the MIT Indoor 67 dataset.

it can be observed that, no matter multi-scale or a single scale level, CTV-GS are always neck and neck with CTV-VB on both datasets. In the case of three individual scales levels, the average difference between CTV-VB and CTV-GS is only 0.70% on the MIT Indoor 67 dataset and 1.38% on the SUN 397 dataset. It decreases to 0.52% between multi-scale CTV-VB and multi-scale CTV-GS on the MIT Indoor 67 dataset and 0.06% on the SUN 397 dataset.

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 patches and involving the logistic normal prior distribution, this method could better model the generative process for feature learning. Implemented on rich semantic information of CNN features, we explore underlying correlated semantics and encode them into the Fisher Vector strategy to increase the discriminative capability. To make the method suiting for the process of large-scale datasets, we further provide Variational Bayesian solution and Gibbs sampling solution. The proposed CTV can be treated as an evolution oriented from Fisher Vector and 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 A
Derivation of CTV with Variational Bayesian solution

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

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

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

where $\{\lambda_i, \nu_i, \phi_{n,i}\}$ are variational parameters as pointed in Section III, and ν is a new introduced variational parameter.

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_{GS} are:

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

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

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

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

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

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, Jul. 2003.

[2] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems*, May 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, Aug. 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 and Systems for Video Technol.*, vol. 13, no. 1, pp. 26–38, Feb. 2003.

[5] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Content-based hierarchical classification of vacation images," in *Proc. IEEE Int. Conf. Multi. 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. Patt. Recog. Workshops*, Jun. 2005, pp. 1–7.

[7] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous 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. Europ. Conf. Comput. Vis.*, May 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. Europ. Conf. Comput. Vis.*, May 2006, pp. 517–530.

[14] A. Bosch, A. Zisserman, and X. Muoz, "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. 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 Int. ACM SIGIR Conf. Res. Devel. Infor. Retriev.*, Jul. 1999, pp. 50–57.

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

[19] D. Blei and J. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, May 2006, pp. 147–154.

[20] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, pp. 17–35, 2007.

[21] J. Aitchison, "The statistical analysis of compositional data," *Monographs on Statistics and Applied Probability Show All Parts in This Series*, Feb. 1986.

[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, Sept. 2007.

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, May 2014, pp. 487–495.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, May 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, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learning*, Jul. 2014, pp. 647–655.

[27] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*, Apr. 1998.

[28] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop on Statistical Learning in Europ. Conf. Comput. Vis.*, Mar. 2004.

[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 *Advances in Neural Information Processing Systems*, May 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*, pp. 1–20, Aug. 2014.

[31] T. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in Neural Information Processing Systems*, pp. 487–493, May 1999.

[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.*, 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 10th 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.*, 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.*, [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7286858, accessed Feb. 24, 2015.

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learning 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, 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, May 2001.

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

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

[43] 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.

[44] J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang, "Scalable inference for logistic-normal topic models," in *Advances in Neural Information Processing Systems*, May 2013, pp. 2445–2453.

[45] D. Mimno, H. M. Wallach, and A. McCallum, "Gibbs sampling for logistic normal topic models with graph-based priors," 2008.

[46] T. Salimans, D. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *Proc. 32nd Int. Conf. Mach. Learning*, no. 37, Jul. 2015, pp. 1218–1226.

[47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.

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

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

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

[51] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Systems. Tech.*, vol. 2, no. 3, p. 27, May 2011.

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

[53] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, N. Vasconcelos, W. Li, and N. Vasconcelos, "Scene classification with semantic fisher vectors," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2974–2983.

[54] 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, Mar. 2014.

[55] 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 *Advances in Neural Information Processing Systems*, May 2014, pp. 1143–1151.
[56] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 923–930.
[57] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Advances in Neural Information Processing Systems*, May 2013, pp. 494–502.