Integrating Image Clustering and Codebook Learning

Pengtao Xie and Eric Xing
{pengtaox,epxing}@cs.cmu.edu
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Image clustering and visual codebook learning are two fundamental problems in computer vision and they are tightly related. On one hand, a good codebook can generate effective feature representations which largely affect clustering performance. On the other hand, class labels obtained from image clustering can serve as supervised information to guide codebook learning. Traditionally, these two processes are conducted separately and their correlation is generally ignored. In this paper, we propose a Double Layer Gaussian Mixture Model (DLGMM) to simultaneously perform image clustering and codebook learning. In DLGMM, two tasks are seamlessly coupled and can mutually promote each other. Cluster labels and codebook are jointly estimated to achieve the overall best performance. To incorporate the spatial coherence between neighboring visual patches, we propose a Spatially Coherent DLGMM which uses a Markov Random Field to encourage neighboring patches to share the same visual word label. We use variational inference to approximate the posterior of latent variables and learn model parameters. Experiments on two datasets demonstrate the effectiveness of two models.

Introduction

Image clustering (Barnard, Duygulu, and Forsyth 2001; Gordon, Greenspan, and Goldberger 2003; Ci et al. 2006; Gao et al. 2005; He et al. 2005; Rege, Dong, and Hua 2008; Aly et al. 2009; Yang et al. 2010) represents a fundamental problem in computer vision and has wide applications in image collection summarization, browsing and analysis. Probably, the most widely used image clustering technique is bag-of-words representation (Sivic and Zisserman 2003; Fei-Fei and Perona 2005; Lazebnik, Schmid, and Ponce 2007) plus K-means clustering (Lloyd 1982), which first converts images into bag-of-words histograms using a learned codebook, then uses K-means method to obtain clusters. Bag-of-words (BOW) model (Sivic and Zisserman 2003; Fei-Fei and Perona 2005; Lazebnik, Schmid, and Ponce 2007) extracts local features (e.g., patches) from images, quantizes their descriptors into visual words based on a visual codebook and uses the resultant histogram of words for downstream tasks such as image clustering, classification and retrieval. To obtain a codebook, visual features extracted from the entire training collection are grouped into clusters and each cluster center is deemed as a visual word and assigned a unique word index.

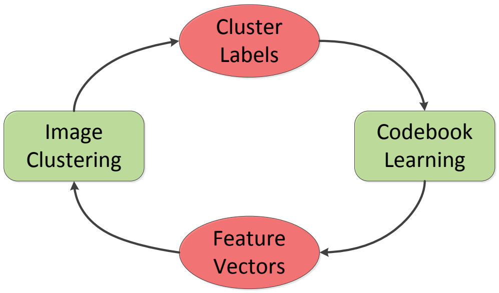

Figure 1: Image clustering and codebook learning are closely related and can mutually promote each other. First, good codebook will produce good feature vectors, which determine the performance of clustering. Second, the cluster labels generated from clustering algorithm can supervise codebook learning. For example, given the information that image A and B are grouped into cluster 1 and image C and D are grouped into cluster 2, a codebook can be learned to make the feature vectors of A and B to be similar and those of A and C to be dissimilar.

Traditionally, codebook learning and image clustering are performed separately. A codebook is first built off-line and images are converted into BOW histograms based on the codebook. Subsequently, clustering is performed over the BOW histograms. This separation ignores the correlation between two tasks. As shown in Figure 1, image clustering and codebook learning are closely coupled and can mutually benefit each other. On one hand, a good codebook can generate effective BOW representations, which are the input of clustering algorithms and largely affect clustering performance. On the other hand, cluster labels obtained from clustering methods can serve as supervised information to guide codebook learning. For example, if knowing two images are likely to be assigned to the same cluster, we can learn a codebook based on which BOW representations of the two images are similar. Clustering and codebook learning follow a chicken-and-egg relationship. Better clustering results produce better codebook and better codebook in turn

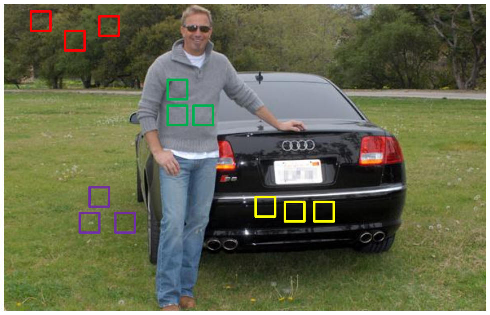

Figure 2: An example showing spatial coherence of image content. Patches marked with red, purple, green, yellow should be mapped to “tree”, “grass”, “sweater” and “car” visual words respectively.

contributes to better clustering results. Performing them separately fails to make them mutually promote each other to achieve the overall best performance. In this paper, we propose a Double Layer Gaussian Mixture Model (DLGMM) to integrate clustering and codebook learning into a unified framework where cluster labels and codebook are jointly estimated. Our model seamlessly couples two layers of Gaussian Mixture Models (GMM). GMM in the first layer is defined over the whole image collection and is used for image clustering. GMM in the second layer is defined over each image and is used for modeling the local patches.

Existing codebook learning methods generally treat local patches as independent and ignore their spatial relationships. Spatial coherence is a salient characteristic of image. An image is composed of a set of non-overlapping scenes and objects. Patches within a scene region or an object usually exhibit strong visual or semantic correlation. For instance, in Figure 2, patches within a certain semantic region, say car, sweater, face, jean, tree, grass, are quite homogeneous. Thereby, when quantizing local patches, it is desirable to assign neighboring patches to the same visual word. As shown in Figure 2, patches marked with red, purple, green, yellow should be mapped to “tree”, “grass”, “sweater” and “car” words respectively. To incorporate the spatial coherence between local patches, we propose a Spatially Coherent Double Layer Gaussian Mixture Model (SC-DLGMM) which uses a Markov Random Field (MRF) (Zhao, Fei-Fei, and Xing 2010) model to encourage nearby patches in an image to share the same visual word label.

The major contributions of our paper are summarized as follows

- We propose a Double Layer Gaussian Mixture Model to perform image clustering and codebook learning simultaneously. Experimental results show that the integration can produce a more effective codebook, which in turn improves clustering performance.
- We propose a Spatially Coherent DLGMM model which incorporates the spatial coherence between neighboring patches in codebook training. Experimental results demonstrate that encoding the spatial correlation of nearby patches can improve the codebook and BOW representations.
- We derive efficient variational inference methods to approximate the posteriors and learn model parameters for the two models.

The rest of this paper is organized as follows. Section 2 reviews related work. In Section 3, we propose the DLGMM model. Section 4 presents SC-DLGMM model. Section 5 gives experimental results. In Section 6, we conclude the paper.

Related Works

Image clustering has been widely studied in (Barnard, Duygulu, and Forsyth 2001; Gordon, Greenspan, and Goldberger 2003; Ci et al. 2006; Gao et al. 2005; He et al. 2005; Rege, Dong, and Hua 2008; Aly et al. 2009; Yang et al. 2010). The most common approach (Gordon, Greenspan, and Goldberger 2003; He et al. 2005; Aly et al. 2009; Yang et al. 2010) is to first represent images into feature vectors, then perform clustering on feature representations. The interconnection between feature learning and clustering are generally ignored. Another line of research (Barnard, Duygulu, and Forsyth 2001; Ci et al. 2006; Gao et al. 2005; He et al. 2005; Rege, Dong, and Hua 2008) focuses on web image clustering. In addition to image contents, these methods utilize textual, link, and meta information to aid clustering, which is beyond the scope of this paper.

Training task-specific codebook (Mairal et al. 2009; Lian et al. 2010; Yang, Yu, and Huang 2010; Fernando et al. 2012; Yang and Yang 2012) has aroused extensive research interests. Supervised codebook learning (Mairal et al. 2009; Lian et al. 2010; Yang, Yu, and Huang 2010; Fernando et al. 2012; Yang and Yang 2012) jointly performs codebook learning and supervised tasks to make the trained codebook optimal for those tasks. Different from their works, our model exploits codebook learning under the context of clustering, which is unsupervised.

DLGMM can be seen as a model jointly modeling observed data and their latent cluster labels. Several topic models (Wang, Ma, and Grimson 2007; Wallach 2008; Zhu et al. 2010; Xie and Xing 2013) have been proposed in this paradigm. These models assume data points inherently belong to several latent clusters and each cluster owns a Dirichlet prior or a Logistic-Normal prior to generate topic proportion vectors for data in this cluster. In these models, each data instance is treated as a combination of topics which are multinomial distributions over textual or visual words. In vision topic models (Wang, Ma, and Grimson 2007; Zhu et al. 2010), codebook is built off-line and each local patch is mapped to a visual word. Then these visual words are modeled using mixture of multinomials. Different from (Wang, Ma, and Grimson 2007; Zhu et al. 2010), our models directly model the descriptors of local patches using Gaussian Mixture Model with the goal of learning a codebook on-line.

SC-DLGMM model borrows the idea of using MRF to encode spatial coherence of local patches from (Verbeek and Triggs 2007; Zhao, Fei-Fei, and Xing 2010) which embed MRF into topic models to encourage neighboring patches to share the same topic label. In their works, spatial coher-

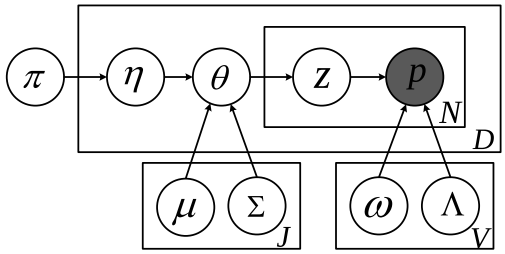

Figure 3: Double Layer Gaussian Mixture Model

ence is imposed over topic labels and they neglect the coherence issue in codebook learning. In our model SC-DLGMM, MRF is defined over visual word labels to encourage nearby patches to be assigned to the same visual word.

Double Layer Gaussian Mixture Model

In this section, we propose a Double Layer Gaussian Mixture Model (DLGMM) and present a variational inference method to approximate the posteriors and learn model parameters.

Model

We assume images are generated from a mixture of clusters where each cluster is associated with a Gaussian distribution over image representations, and assume visual patches are generated from a mixture of visual words where each visual word is modeled with a Gaussian distribution over visual descriptors. Based on these assumptions, we propose a DLGMM model (Figure 3), which seamlessly couples two layers of Gaussian Mixture Models (GMM). GMM in the first layer is composed of $\boldsymbol{\pi}$, η, $\boldsymbol{\theta}$, $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^J$, which is defined over the whole image collection and is used for image clustering. GMM in the second layer is composed of $\boldsymbol{\theta}$, z, p, $\{\boldsymbol{\omega}_v, \boldsymbol{\Lambda}_v\}_{v=1}^V$, which is defined over each image and is used for modeling the visual patches. θ is the latent representation of an image, which ties the two layers of GMMs together to bridge image clustering and codebook learning. θ is the observation of the first-layer GMM and acts as the mixture weights of the second-layer GMM.

Given an image collection containing D images, we assume these images inherently belong to J groups. We assume there exists a codebook containing V visual words and each visual word has a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\omega}, \boldsymbol{\Lambda})$ over visual patch descriptors. For simplicity, we assume covariance matrix is isotropic, $\boldsymbol{\Lambda} = \delta^2\mathbf{I}$. Each group has a group-specific Logistic-Normal prior $\mathcal{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which is used for sampling multinomial distributions over visual words. The Logistic-Normal is a distribution on the simplex that allows for a general pattern of variability between the components by transforming a multivariate Gaussian random variable (Blei and Lafferty 2006; Ahmed and Xing 2007). The multivariate Gaussian random variable $\boldsymbol{\theta}$ of documents in group j are sampled from $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ and are converted to multinomial distributions using Logistic mapping. Another commonly used prior for multinomials is Dirichlet distribution (Blei, Ng, and Jordan 2003). The reason to choose Logistic-Normal prior rather than Dirichlet prior is to capture the correlation (Blei and Lafferty 2006; Ahmed and Xing 2007) between visual words through the covariance matrix $\boldsymbol{\Sigma}$. There usually exists strong correlation between visual words. For example, a "sky" visual word is more likely to co-occur with a "sun" word than a "car" visual word. Dirichlet prior is unable to model these correlations. A global multinomial prior $\boldsymbol{\pi}$ is used to choose group membership for an image. π_j denotes the prior probability that an image belongs to group j.

Each image is associated with a group indicator and has a multivariate Gaussian random variable to generate visual word labels. Visual patches in an image are generated from visual words. To generate an image containing N visual patches $\mathbf{p} = \{p_i\}_{i=1}^N$, we first choose a group[1] η from the multinomial distribution parametrized by $\boldsymbol{\pi}$. Then from the Gaussian prior $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ corresponding to group η, we sample a Gaussian variable $\boldsymbol{\theta}$ and map $\boldsymbol{\theta}$ to a simplex using Logistic function. To generate a patch p, we first pick up a visual word[2] z from $\boldsymbol{\theta}$

$$p(z|\boldsymbol{\theta}) = \frac{\prod_{v=1}^{V}[\exp(\theta_v)]^{z_v}}{\sum_{l=1}^{V}\exp(\theta_l)} \tag{1}$$

then generate the descriptor $\mathbf{o}$ of this patch from the multivariate Gaussian distribution corresponding to visual word z.

The generative process of an image in DLGMM can be summarized as follows

- Sample a group $\eta \sim Multinomial(\boldsymbol{\pi})$
- Sample $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\eta, \boldsymbol{\Sigma}_\eta)$
- For each patch p
  - sample a visual word z according to Eq.(1)
  - sample patch descriptor $\mathbf{o} \sim \mathcal{N}(\boldsymbol{\omega}_z, \boldsymbol{\Lambda}_z)$

Accordingly, the joint distribution of η, $\boldsymbol{\theta}$, $\mathbf{z} = \{z_i\}_{i=1}^N$, $\mathbf{O} = \{\mathbf{o}_i\}_{i=1}^N$ given model parameters $\boldsymbol{\pi}$, $\mathbf{G}_1 = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^J$, $\mathbf{G}_2 = \{\boldsymbol{\omega}_v, \boldsymbol{\Lambda}_v\}_{v=1}^V$ can be written as

$$\begin{aligned} & p(\eta,\boldsymbol{\theta},\mathbf{z},\mathbf{O}|\boldsymbol{\pi},\mathbf{G}_1,\mathbf{G}_2) \\ &= p(\eta|\boldsymbol{\pi})p(\boldsymbol{\theta}|\eta,\mathbf{G}_1)p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{O}|\mathbf{z},\mathbf{G}_2)\\ &= \prod_{j=1}^J \pi_j^{\eta_j}\prod_{j=1}^J[\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)]^{\eta_j}\\ & \prod_{i=1}^N \frac{\prod_{v=1}^V[\exp(\theta_v)]^{z_{iv}}}{\sum_{l=1}^V\exp(\theta_l)}\prod_{v=1}^V[\mathcal{N}(\mathbf{o}_i|\boldsymbol{\omega}_v,\boldsymbol{\Lambda}_v)]^{z_{iv}} \end{aligned} \tag{2}$$

We believe that performing image clustering and codebook learning jointly is superior to doing them separately. As stated above, in DLGMM, image clustering is accomplished by estimating parameters of GMM in the first layer and codebook learning involves estimating parameters of GMMs in the second layer. Performing clustering and codebook learning separately is equivalent to estimating parameters of GMM in one layer while fixing those in the other

[1] η is a 1-of-J vector of size J with one component equals to 1.

[2] z is a 1-of-V vector (size V) with one component equals to 1.

layer. In the case where we first build a codebook off-line based on which images are represented with BOW histograms and then perform clustering, we are actually clamping parameters of GMMs in the second layer to some predefined values and then estimating those in the first layer. In the other case where codebook learning follows clustering, parameters of GMM in the first layer are predefined and we estimate those in the second layer. In contrast, performing the two tasks jointly is equivalent to estimating parameters of GMMs in two layers simultaneously.

Variational Inference and Parameter Learning

The key inference problem involved in DLGMM is to estimate the posterior distribution $p(\eta, \boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\pi}, \mathbf{G}_1, \mathbf{G}_2)$ of latent variables $\mathbf{H} = \{\eta, \boldsymbol{\theta}, \mathbf{z}\}$ given observed variables $\mathbf{O}$ and model parameters $\boldsymbol{\Pi} = \{\boldsymbol{\pi}, \mathbf{G}_1, \mathbf{G}_2\}$. Since exact inference is intractable, we use variational inference (Wainwright and Jordan 2008) to approximate the posterior.

The variational distribution q is defined as follows

$$q(\eta, \boldsymbol{\theta}, \mathbf{z}) = q(\eta|\boldsymbol{\zeta}) q(\boldsymbol{\theta}|\boldsymbol{\alpha}, \tau^2\mathbf{I}) \prod_{i=1}^{N} q(z_i|\boldsymbol{\phi}_i) \tag{3}$$

where $\boldsymbol{\zeta}$ and $\{\boldsymbol{\phi}_i\}_{i=1}^N$ are multinomial parameters. $\boldsymbol{\alpha}$ and $\mathrm{diag}(\boldsymbol{\tau}^2)\mathbf{I}$ are mean and covariance of Gaussian distribution. Given the variational distribution, we can derive a variational lower bound, which can be optimized using an EM algorithm.

In E-step, we update variational parameters as follows

$$\begin{aligned}\zeta_j \propto \pi_j \exp\{&-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_j| - \tfrac{1}{2}\mathrm{tr}(\mathrm{diag}(\boldsymbol{\tau}^2)\boldsymbol{\Sigma}_j^{-1}) \\ &-\tfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\mu}_j)^{\mathsf{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}_j)\}\end{aligned} \tag{4}$$

$$\phi_{iv} \propto \exp\{\alpha_v - \frac{R}{2}\log\delta_v^2 - \frac{(\mathbf{o}_i - \boldsymbol{\omega}_v)^{\mathsf{T}}(\mathbf{o}_i - \boldsymbol{\omega}_v)}{2\delta_v^2})\} \tag{5}$$

where R is the dimension of image descriptor.

$$e = \sum_{l=1}^{V} \exp\{\alpha_l + \frac{\tau_l^2}{2}\} \tag{6}$$

where e is a newly introduced variational variable. The analytical maximization w.r.t $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}^2$ is not amenable. Instead, we use gradient descent method to optimization these two variables.

In M-step, we update model parameters by maximizing the lower bound defined over a set of images $\{\mathbf{O}_d\}_{d=1}^D$

$$\pi_j = \frac{\sum_{d=1}^{D}\zeta_{dj}}{D}, \boldsymbol{\mu}_j = \frac{\sum_{d=1}^{D}\zeta_{dj}\boldsymbol{\alpha}_d}{\sum_{d=1}^{D}\zeta_{dj}} \tag{7}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{d=1}^{D}\zeta_{dj}(\boldsymbol{\tau}_d^2\mathbf{I} + (\boldsymbol{\alpha}_d - \boldsymbol{\mu}_j)(\boldsymbol{\alpha}_d - \boldsymbol{\mu}_j)^{\mathsf{T}})}{\sum_{d=1}^{D}\zeta_{dj}} \tag{8}$$

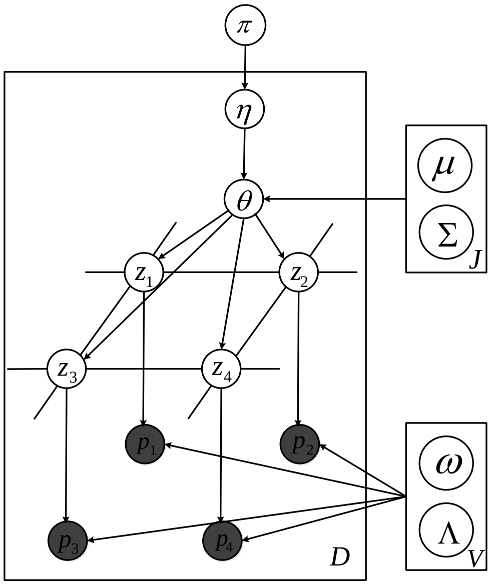

Figure 4: Spatially Coherent Double Layer Gaussian Mixture Model

$$\boldsymbol{\omega}_v = \frac{\sum_{d=1}^{D}\sum_{i=1}^{N_d}\phi_{d,i,v}\mathbf{o}_{di}}{\sum_{d=1}^{D}\sum_{i=1}^{N_d}\phi_{d,i,v}} \tag{9}$$

$$\delta_v^2 = \frac{\sum_{d=1}^{D}\sum_{i=1}^{N_d}\phi_{d,i,v}(\mathbf{o}_{di} - \boldsymbol{\omega}_v)^{\mathsf{T}}(\mathbf{o}_{di} - \boldsymbol{\omega}_v)}{R\sum_{d=1}^{D}\sum_{i=1}^{N_d}\phi_{d,i,v}} \tag{10}$$

Spatially Coherent Double Layer Gaussian Mixture Model

In DLGMM model, the visual word labels for patches are independently assigned, which falsely ignores the spatial relationships between neighboring patches. As a remedy, we propose a Spatially Coherent Double Layer Gaussian Mixture Model (SC-DLGMM) model, which uses Markov Random Field (MRF) model to ensure spatial coherence in visual word assignments.

Model

As shown in Figure 4, we define a Markov Random Field on the latent visual word layer to encourage neighboring patches to share the same visual word label. Specifically, we define the joint distribution of visual word assignments $\mathbf{z} = \{z_i\}_{i=1}^N$ for all patches in an image as

$$p(\mathbf{z}|\boldsymbol{\theta}, \gamma) = \frac{1}{Z(\boldsymbol{\theta}, \gamma)}\prod_{i=1}^{N} p(z_i|\boldsymbol{\theta}) \exp\{\gamma \sum_{(m,n)\in\mathcal{P}} \mathbb{I}(z_m = z_n)\} \tag{11}$$

where $Z(\boldsymbol{\theta}, \gamma)$ denotes the partition function

$$Z(\boldsymbol{\theta}, \gamma) = \sum_{\boldsymbol{z}}\prod_{i=1}^{N} p(z_i|\boldsymbol{\theta}) \exp\{\gamma \sum_{(m,n)\in\mathcal{P}} \mathbb{I}(z_m = z_n)\} \tag{12}$$

Table 1: Clustering accuracy (%) on 15-Scenes dataset

Codebook Size	100	200	300	400	500	600	700	800	900	1000
KM	26.98	27.71	30.26	28.65	29.68	29.79	28.87	28.78	29.05	29.88
NC	25.17	27.45	27.42	26.64	25.66	26.31	26.56	26.33	28.76	28.52
JSOM	27.34	27.11	28.97	29.45	28.98	27.33	27.56	28.19	30.08	27.22
LDA	34.11	33.02	27.58	34.72	31.06	31.84	**36.70**	**35.14**	31.88	29.81
DLGMM	**34.74**	**35.03**	34.74	34.78	34.45	35.28	34.47	34.18	34.02	34.18
SC-DLGMM	34.16	34.20	**34.95**	**35.23**	**34.81**	**35.57**	34.29	34.85	**34.61**	**34.27**

Table 2: Clustering accuracy (%) on Caltech-101 dataset

Codebook Size	100	200	300	400	500	600	700	800	900	1000
KM	13.31	13.46	14.27	14.33	14.57	14.60	14.55	14.21	13.79	13.98
NC	14.01	14.01	13.81	13.92	13.79	14.14	14.22	14.08	13.60	13.55
JSOM	12.87	12.98	13.00	12.76	12.89	12.57	12.96	13.07	12.76	12.33
LDA	13.99	17.09	18.30	18.95	17.05	19.08	18.38	19.39	18.23	18.91
DLGMM	**21.42**	20.94	20.87	20.46	20.92	20.02	20.07	20.09	20.31	20.17
SC-DLGMM	21.27	**21.20**	**20.98**	**20.86**	**21.06**	**20.42**	**20.42**	**20.39**	**20.44**	**20.48**

$\mathbb{I}(\cdot)$ denotes the indicator function and $\mathcal{P}$ denotes all connected pairs of patches. A positive value of γ awards configurations where neighboring patches share the same word label. $p(z_i|\boldsymbol{\theta})$ is defined the same as that in Eq.(1).

The generative process of an image in SC-DLGMM can be summarized as follows

- Sample a group $\eta \sim Multinomial(\boldsymbol{\pi})$
- Sample $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\eta, \boldsymbol{\Sigma}_\eta)$
- Sample $\mathbf{z}$ jointly for all patches using Eq.(11)
- For each patch p, sample $\mathbf{o} \sim \mathcal{N}(\boldsymbol{\omega}_z, \boldsymbol{\Lambda}_z)$

Accordingly, the joint distribution of η, $\boldsymbol{\theta}$, $\mathbf{z}$, $\mathbf{O}$ can be written as

$$\begin{aligned} & p(\eta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{O}|\boldsymbol{\pi}, \mathbf{G}_1, \mathbf{G}_2, \gamma) \\ & = p(\eta|\boldsymbol{\pi})p(\boldsymbol{\theta}|\eta, \mathbf{G}_1)p(\mathbf{z}|\boldsymbol{\theta}, \gamma)p(\mathbf{O}|\mathbf{z}, \mathbf{G}_2) \end{aligned} \tag{13}$$

Variational Inference and Parameter Learning

We use variational inference method to approximate posteriors and estimate model parameters. The variational distribution q is the same as that defined in Eq.(3).

The updates of ζ_j, e, $\boldsymbol{\alpha}$, $\boldsymbol{\tau}^2$, π_j, $\boldsymbol{\mu}_j$, σ_j^2, $\boldsymbol{\omega}_v$, δ_v^2 are the same as those in DLGMM. Variational parameter ϕ_{iv} can be computed as

$$\begin{aligned} \phi_{iv} \;\; \propto \;\; & \exp\{\alpha_v - \tfrac{R}{2}\log\delta_v^2 - \tfrac{(\mathbf{o}_i-\boldsymbol{\omega}_v)^{\mathsf{T}}(\mathbf{o}_i-\boldsymbol{\omega}_v)}{2\delta_v^2}) \\ + \;\; & \gamma \sum_{n\in\mathcal{N}(i)} \phi_{nv}\} \end{aligned} \tag{14}$$

where $\mathcal{N}(i)$ is the patches connected with patch i. ϕ_{iv} indicates how likely patch i will be assigned to word v. From Eq.(14), we can see that the update of ϕ_{iv} of patch i depends on the ϕ_{nv} of i's neighbors n. This mechanism imposes spatial consistency. The tradeoff parameter γ is hard to learn in that it cannot be updated in closed form in each iteration. Hence, we choose to hand-tune it.

Experiments

In this section, we evaluate the effectiveness of DLGMM and SC-DLGMM models by comparing them with four baseline methods on image clustering task.

Experimental Settings

The experiments are conducted on 15-Scenes (Lazebnik, Schmid, and Ponce 2007) dataset and Caltech-101 (Fei-Fei, Fergus, and Perona 2004) dataset. The 15-Scenes dataset contains 4485 images which are grouped into 15 scene categories. Caltech-101 dataset contains 9144 images from 101 object categories, from which we randomly choose half images for our experiments. Following (Lazebnik, Schmid, and Ponce 2007), we densely extract local patches of size 16×16 on a grid with stepsize 16. Each patch is represented with SIFT (Lowe 2004) descriptor whose dimensionality is 128. We collect about 11M patches from 15-Scenes dataset and about 13M patches from Caltech-101 dataset.

We use two metrics to measure the clustering performance: accuracy (AC) and normalized mutual information (NMI). Please refer to (Cai, He, and Han 2011) for detailed definition of these two metrics. We compare our models with four methods: K-means (KM), Normalized Cut (NC) (Shi and Malik 2000), joint scene object model (JSOM) (Zhu et al. 2010) and Latent Dirichlet Allocation (LDA). K-means and Normalized Cut are probably the most widely used clustering algorithms. Like our models, JSOM simultaneously performs image clustering and modeling. The key difference is JSOM first quantizes local patches into visual words using a pre-trained codebook and subsequently uses mixture of multinomials to model visual words. Our models use mixture of Gaussians to model local patches and the codebook is learned on-line. LDA (Lu, Mei, and Zhai 2011) can be used for clustering by treating each topic as a cluster. An image is assigned to cluster x if $x = \mathrm{argmax}_k \theta_k$, where $\boldsymbol{\theta}$ is the topic proportion vector of the image. For these four baseline methods, we use K-means to train the codebook on

Table 3: Normalized mutual information (%) on 15-Scenes dataset

Codebook Size	100	200	300	400	500	600	700	800	900	1000
KM	25.52	26.79	27.42	26.47	28.64	28.38	28.72	27.84	28.84	28.69
NC	23.16	24.58	24.16	25.65	24.56	24.39	25.47	24.92	26.61	26.88
JSOM	26.86	26.89	27.92	28.56	27.85	27.02	27.45	28.09	28.47	26.98
LDA	31.50	30.39	28.94	**32.94**	30.38	30.57	**32.21**	**34.43**	31.48	30.57
DLGMM	**32.23**	**32.81**	32.29	32.30	32.21	32.13	32.10	32.15	32.09	31.98
SC-DLGMM	31.53	32.32	**32.52**	32.44	**32.67**	**32.51**	32.09	32.37	**32.55**	**32.29**

Table 4: Normalized mutual information (%) on Caltech-101 dataset

Codebook Size	100	200	300	400	500	600	700	800	900	1000
KM	37.02	37.11	37.24	37.33	37.46	**37.41**	**37.35**	**37.07**	**37.01**	**37.05**
NC	35.85	35.33	35.11	35.26	35.29	35.44	35.48	35.32	35.28	35.36
JSOM	33.67	34.72	34.94	34.64	34.99	34.59	34.80	35.20	34.72	34.18
LDA	32.98	35.46	36.18	35.92	35.35	36.59	36.23	36.50	35.70	36.56
DLGMM	38.50	38.14	37.44	37.28	36.38	36.91	36.74	36.72	36.58	36.40
SC-DLGMM	**38.60**	**38.55**	**37.53**	**37.50**	**37.69**	37.16	36.97	36.92	36.74	36.56

all collected image patches and obtain bag-of-words (BOW) representations of images using vector quantization. BOW vectors are weighted using tf-idf and are normalized to unit length using L2 norm. The required input cluster number in KM, NC and our models is set to the ground truth number of categories in datasets. In NC, we use Gaussian kernel as similarity measure between images. The bandwidth parameter is set to 1. In JSOM, topic number is set to 100. In LDA, symmetric Dirichlet priors are used and are set to 0.05. In SC-DLGMM, parameter γ on the MRF is tuned to produce the best possible clustering performance. Our models are initialized with the clustering results obtained from LDA. We compare these methods under varying codebook size ranging from 100 to 1000 with an increment of 100. JSOM and SC-DLGMM are probabilistic models where each image has a distribution over clusters. We assign each image to the most probable cluster.

Results

Table 1 and 2 summarize the clustering accuracy on 15-Scenes dataset and Caltech-101 dataset. Table 3 and 4 summarize the normalized mutual information on 15-Scenes dataset and Caltech-101 dataset. As can be seen from the results, our models DLGMM and SC-DLGMM are superior to the three baseline methods on both datasets and both evaluation metrics. This corroborates our assumption that performing clustering and codebook learning jointly can achieve better performance than doing than separately. In baseline methods, codebook is first learned off-line and clustering is conducted subsequently on the image feature vectors built from the codebook. Usually, the codebook is learned with K-means algorithm or Gaussian mixture model, with the goal to maximize the likelihood of image patches. A codebook learned in such way is irrelevant to any specific higher level tasks, including clustering, classification and retrieval. When applied to clustering, the codebook is not guaranteed to deliver desirable clustering performance. DLGMM and SC-DLGMM combine codebook learning and clustering into a unified framework where the two tasks are jointly performed. In each iteration of the inference and learning process, the cluster assignments of images depend on the current learned codebook and the estimation of visual words depends on the current inferred cluster labels. The learning of codebook is continually guided by intermediate clustering results, thereby it is specifically suitable for clustering task in the end.

Comparing DLGMM and SC-DLGMM, we can see that SC-DLGMM further improve the clustering performance. SC-DLGMM incorporates the spatial coherence of neighboring pixels and defines a MRF over the latent word assignments layer to encourage neighboring pixels to share the same word label. DLGMM ignores the relationships between pixels and each pixel is tackled independently. Thereby, DLGMM is inferior to SC-DLGMM.

Conclusions

We study the problem of jointly image clustering and codebook learning and propose two models: DLGMM and SC-DLGMM. In DLGMM, image clustering and codebook learning are integrated into a unified framework to make two tasks mutually benefit each other. In SC-DLGMM, we investigate the spatial coherence of image content and encourage neighboring patches to share the same visual word. Experiments on two datasets demonstrate that: 1, integrating image clustering and codebook learning can produce a better codebook; 2, incorporating spatial coherence between neighboring patches can improve the effectiveness of codebook.

Acknowledgments

We would like to thank the anonymous reviewers and Shanghang Zhang for their valuable comments and suggestions. This work was supported by NSF IIS1111142, NSF IIS1447676 and AFOSR FA95501010247.

References

Ahmed, A., and Xing, E. P. 2007. On tight approximate inference of logistic-normal admixture model. In *In Proceedings of the Eleventh International Conference on Artifical Intelligence and Statistics*. Citeseer.

Aly, M.; Welinder, P.; Munich, M.; and Perona, P. 2009. Towards automated large scale discovery of image families. In *Computer Vision and Pattern Recognition Workshops*.

Barnard, K.; Duygulu, P.; and Forsyth, D. 2001. Clustering art. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Blei, D., and Lafferty, J. 2006. Correlated topic models. *Advances in neural information processing systems*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3:993–1022.

Cai, D.; He, X.; and Han, J. 2011. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Ci, D.; He, X.; Li, Z.; Ma, W.-Y.; and Wen, J.-R. 2006. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th annual ACM international conference on Multimedia*.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop*.

Fernando, B.; Fromont, E.; Muselet, D.; and Sebban, M. 2012. Supervised learning of gaussian mixture models for visual vocabulary generation. *Pattern Recognition*.

Gao, B.; Liu, T.-Y.; Qin, T.; Zheng, X.; Cheng, Q.-S.; and Ma, W.-Y. 2005. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th annual ACM international conference on Multimedia*.

Gordon, S.; Greenspan, H.; and Goldberger, J. 2003. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In *International Conference on Computer Vision*.

He, X.; Cai, D.; Liu, H.; and Han, J. 2005. Image clustering with tensor representation. In *Proceedings of the 13th annual ACM international conference on Multimedia*.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2007. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lian, X.-C.; Li, Z.; Lu, B.-L.; and Zhang, L. 2010. Max-margin dictionary learning for multiclass image categorization. *European Conference on Computer Vision*.

Lloyd, S. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*.

Lu, Y.; Mei, Q.; and Zhai, C. 2011. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval* 14(2):178–203.

Mairal, J.; Ponce, J.; Sapiro, G.; Zisserman, A.; and Bach, F. R. 2009. Supervised dictionary learning. In *Advances in neural information processing systems*, 1033–1040.

Rege, M.; Dong, M.; and Hua, J. 2008. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In *Proceedings of the 17th international conference on World Wide Web*.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sivic, J., and Zisserman, A. 2003. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*.

Verbeek, J., and Triggs, B. 2007. Region classification with markov field aspect models. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*.

Wallach, H. M. 2008. Structured topic models for language. *Unpublished doctoral dissertation, Univ. of Cambridge*.

Wang, X.; Ma, X.; and Grimson, E. 2007. Unsupervised activity perception by hierarchical bayesian models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Xie, P., and Xing, E. P. 2013. Integrating document clustering and topic modeling. In *Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence*.

Yang, J., and Yang, M.-H. 2012. Top-down visual saliency via joint crf and dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, Y.; Xu, D.; Nie, F.; Yan, S.; and Zhuang, Y. 2010. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*.

Yang, J.; Yu, K.; and Huang, T. 2010. Supervised translation-invariant sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, B.; Fei-Fei, L.; and Xing, E. 2010. Image segmentation with topic random field. *European Conference on Computer Vision*.

Zhu, J.; Li, L.-J.; Fei-Fei, L.; and Xing, E. P. 2010. Large margin learning of upstream scene understanding models. *Advances in Neural Information Processing Systems*.