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.

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 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.
We propose a Spatially Coherent DLGMM model which

• We derive efficient variational inference methods to approx-
  imate the posteriors and learn model parameters for
  the two models.

The rest of this paper is organized as follows. Section 2 re-
views 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 pa-
per.

Related Works

Image clustering has been widely studied in (Barnard,
Duygulu, and Forsyth 2001; Gordon, Greenspan, and Gold-
berger 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 meth-
ods utilize textual, link, and meta information to aid clus-
tering, 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 inter-
est. 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 opti-
mal 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 ob-
served data and their latent cluster labels. Several topic mod-
els (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 be-
long to several latent clusters and each cluster owns a Dirich-
let prior or a Logistic-Normal prior to generate topic propor-
tion 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 Gauss-
ian Mixture Model with the goal of learning a codebook
on-line.

SC-DLGMM model borrows the idea of using MRF to en-
code 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 2: An example showing spatial coherence of im-
age content. Patches marked with red, purple, green, yellow
should be mapped to “tree”, “grass”, “sweater” and “car”
visual words respectively.
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 image representations, and assume visual patches are generated from a mixture of clusters. 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 \( \pi, \eta, \theta, \{ \mu_j, \Sigma_j \}_{j=1}^{J} \), which is defined over the whole image collection and is used for image clustering. GMM in the second layer is composed of \( \theta, \zeta, p, \{ \omega_v, \Lambda_v \}_{v=1}^{V} \), which is defined over each image and is used for modeling the visual patches. \( \theta \) is the latent representation of an image, which ties the two layers of GMMs together to bridge image clustering and codebook learning. \( \theta \) 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 \( \mathcal{N}(\omega, \Lambda) \) over visual patch descriptors. For simplicity, we assume covariance matrix is isotropic, \( \Lambda = \delta I \). Each group has a group-specific Logistic-Normal prior \( \mathcal{LN}(\mu, \Sigma) \) 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 \( \theta \) of documents in group \( j \) are sampled from \( \mathcal{N}(\mu_j, \Sigma_j) \) 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 \( \Sigma \). There usually exists strong correlation between visual words. For example, a “sky” visual word is more likely to co-occur with a “cloud” word rather than a “car” visual word. Dirichlet prior is unable to model these correlations. A global multinomial prior \( \pi \) is used to choose group membership for an image. \( \pi_j \) 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 a mixture of clusters. To generate an image containing \( N \) visual patches \( p = \{ p_i \}_{i=1}^{N} \), we first choose a group \( \eta \) from the multinomial distribution parametrized by \( \pi \). Then from the Gaussian prior \( \mathcal{N}(\mu, \Sigma) \) corresponding to group \( \eta \), we sample a Gaussian variable \( \theta \) and map \( \theta \) to a simplex using Logistic function. To generate a patch \( p \), we first pick up a visual word \( z \) from \( \theta \)

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

then generate the descriptor \( o \) 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 \( \eta \sim \text{Multinomial}(\pi) \)
- Sample \( \theta \sim \mathcal{N}(\mu_\eta, \Sigma_\eta) \)
- For each patch \( p \)
  - sample a visual word \( z \) according to Eq.(1)
  - sample patch descriptor \( o \sim \mathcal{N}(\omega_z, \Lambda_z) \)

Accordingly, the joint distribution of \( \eta, \theta, z = \{ z_i \}_{i=1}^{N} \), \( O = \{ o_{i} \}_{i=1}^{N} \) given model parameters \( \pi \), \( G_1 = \{ \mu_j, \Sigma_j \}_{j=1}^{J} \), \( G_2 = \{ \omega_v, \Lambda_v \}_{v=1}^{V} \) can be written as

\[
\begin{aligned}
p(\eta, \theta, z, O | \pi, G_1, G_2) &= p(\eta | \pi)p(\theta | \eta, G_1)p(z | \theta, G_2) \\
&= \prod_{j=1}^{J} \pi_j^{\eta_j} \prod_{j=1}^{J} [\mathcal{N}(\theta | \mu_j, \Sigma_j)]^{\eta_j} \\
&\quad \times \prod_{i=1}^{N} \frac{\prod_{v=1}^{V} [\exp(\theta_v)]^{z_{iv}}}{\sum_{l=1}^{L} \exp(\theta_l)} \prod_{i=1}^{N} [\mathcal{N}(o_i | \omega_v, \Lambda_v)]^{z_{iv}}
\end{aligned}
\]

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
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 \( p(\eta, \theta, z | \pi, G_1, G_2) \) of latent variables \( H = \{ \eta, \theta, z \} \) given observed variables \( O \) and model parameters \( \Pi = \{ \pi, G_1, G_2 \} \). 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

\[
q(\eta, \theta, z) = q(\eta) q(\theta | \tau^2 I) \prod_{i=1}^{N} q(z_i | \phi_i) \quad (3)
\]

where \( \zeta \) and \( \{ \phi_i \}_{i=1}^{N} \) are multinomial parameters, \( \alpha \) and \( \text{diag}(\tau^2 I) \) 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

\[
\begin{align*}
\zeta_j &\propto \pi_j \exp \left\{ -\frac{1}{2} \log |\Sigma_j| - \frac{1}{2} \text{tr}(\text{diag}(\tau^2 \Sigma_j^{-1})) \right. \\
&\quad -\frac{1}{2}(\alpha - \mu_j)^T \Sigma_j^{-1}(\alpha - \mu_j) \right\} \\
\phi_{iv} &\propto \exp \left\{ \alpha_v - \frac{R}{2} \log \delta_v^2 - \frac{(o_v - \omega_v)^T(o_v - \omega_v)}{2\delta_v^2} \right\} \\
\end{align*}
\]

where \( R \) is the dimension of image descriptor.

\[
e = \sum_{i=1}^{V} \exp \{ \alpha_i + \frac{\tau^2}{2} \} \quad (6)
\]

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

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

\[
\begin{align*}
\pi_j &= \frac{\sum_{d=1}^{D} \zeta_{dj}}{D}, \quad \mu_j = \frac{\sum_{d=1}^{D} \zeta_{dj} \alpha_d}{\sum_{d=1}^{D} \zeta_{dj}} \quad (7) \\
\Sigma_j &= \frac{\sum_{d=1}^{D} \zeta_{dj}(\tau^2 I + (\alpha_d - \mu_j)(\alpha_d - \mu_j)^T)}{\sum_{d=1}^{D} \zeta_{dj}} \\
\end{align*}
\]

**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 \( z = \{ z_{i} \}_{i=1}^{N} \) for all patches in an image as

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

where \( Z(\theta, \gamma) \) denotes the partition function

\[
Z(\theta, \gamma) = \sum_{z} \prod_{i=1}^{N} p(z_i | \theta) \exp \{ \gamma \sum_{(m,n) \in \mathcal{P}} \mathbb{I}(z_m = z_n) \} \\
\]
I(·) denotes the indicator function and $P$ denotes all connected pairs of patches. A positive value of $\gamma$ awards configurations where neighboring patches share the same word label. $p(z_i|\theta)$ is defined as the same as in Eq.(3).

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

- Sample a group $\eta \sim \text{Multinomial}(\pi)$
- Sample $\theta \sim \mathcal{N}(\mu, \Sigma)$
- Sample $z$ jointly for all patches using Eq.(11)
- For each patch $p$, sample $o \sim \mathcal{N}(\omega_z, \Lambda_z)$

Accordingly, the joint distribution of $\eta$, $\theta$, $z$, $O$ can be written as

$$p(\eta, \theta, z, O|\pi, G_1, G_2, \gamma) = p(\eta|\pi)p(\theta|\eta, G_1)p(z|\theta, \gamma)p(O|z, G_2)$$ (13)

**Variational Inference and Parameter Learning**

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

The updates of $\zeta_j$, $\epsilon$, $\alpha^2$, $\omega_j$, $\omega_v$, $\delta^2_v$ are the same as those in DLGMM. Variational parameter $\phi_{iv}$ can be computed as

$$\phi_{iv} \propto \exp\left\{\alpha_v - \frac{1}{2} \log \sigma^2_v - \frac{(\alpha_v - \omega_v)^T(\alpha_v - \omega_v)}{2\sigma^2_v} + \gamma \sum_{n \in \mathcal{N}(i)} \phi_{nv}\right\}$$ (14)

where $\mathcal{N}(i)$ is the patches connected with patch $i$. $\phi_{iv}$ indicates how likely patch $i$ will be assigned to word $v$. From Eq.(14), we can see that the update of $\phi_{iv}$ of patch $i$ depends on the $\phi_{nv}$ of $i$'s neighbors $n$. This mechanism imposes spatial consistency. The tradeoff parameter $\gamma$ 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 $16 \times 16$ on a grid with step size 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 $x = \arg\max_k \theta_k$, where $\theta$ 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

<table>
<thead>
<tr>
<th>Codebook Size</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM</td>
<td>25.52</td>
<td>26.79</td>
<td>27.42</td>
<td>26.47</td>
<td>28.64</td>
<td>28.38</td>
<td>28.72</td>
<td>27.84</td>
<td>28.84</td>
<td>28.69</td>
</tr>
<tr>
<td>JSOM</td>
<td>26.86</td>
<td>26.89</td>
<td>27.92</td>
<td>28.56</td>
<td>27.85</td>
<td>27.02</td>
<td>27.45</td>
<td>28.09</td>
<td>28.47</td>
<td>26.98</td>
</tr>
<tr>
<td>LDA</td>
<td>31.50</td>
<td>31.38</td>
<td>32.94</td>
<td>32.38</td>
<td>32.21</td>
<td>32.13</td>
<td>32.15</td>
<td>32.09</td>
<td>31.98</td>
<td></td>
</tr>
<tr>
<td>DLGMM</td>
<td>32.23</td>
<td>32.81</td>
<td>32.52</td>
<td>32.44</td>
<td>32.67</td>
<td>32.51</td>
<td>32.09</td>
<td>32.37</td>
<td>32.55</td>
<td>32.29</td>
</tr>
<tr>
<td>SC-DLGMM</td>
<td>31.53</td>
<td>32.32</td>
<td>32.52</td>
<td>32.44</td>
<td>32.67</td>
<td>32.51</td>
<td>32.09</td>
<td>32.37</td>
<td>32.55</td>
<td>32.29</td>
</tr>
</tbody>
</table>

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

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

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 so 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 word label. 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