Multimed Tools Appl
DOI 10.1007/s11042-014-1916-3

Learning latent semantic model with visual consistency for image analysis

Jian Cheng · Peng Li · Ting Rui · Hanqing Lu

Abstract Latent semantic models (e.g. PLSA and LDA) have been successfully used in document analysis. In recent years, many of the latent semantic models have also been proved to be promising for visual content analysis tasks, such as image clustering and classification. The topics and words which are two of the key components in latent semantic models have explicit semantic meaning in document analysis. However, these topics and words are difficult to be described or represented in visual content analysis tasks, which usually leads to failure in practice. In this paper, we consider simultaneously the topic consistency and word consistency in semantic space to adapt the traditional PLSA model to the visual content analysis tasks. In our model, the ℓ^1-graph is constructed to model the local neighborhood structure of images in feature space and the word co-occurrence is computed to capture the local word consistency. Then, the local information is incorporated into the model for topic discovering. Finally, the generalized EM algorithm is used to estimate the parameters. Extensive experiments on publicly available databases demonstrate the effectiveness of our approach.

Keywords PLSA · Latent semantic model · Image clustering

J. Cheng (✉) · P. Li · H. Lu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
Beijing 100190, China
e-mail: jcheng@nlpr.ia.ac.cn

P. Li
e-mail: pli@nlpr.ia.ac.cn

H. Lu
e-mail: luhq@nlpr.ia.ac.cn

T. Rui
PLA University of Sciences and Technology, Nanjing 210007, China

Published online: 29 March 2014

1 Introduction

Due to well-known "semantic gap", low-level features cannot fully describe the high-level semantic meaning of images or videos, which has become the main obstacle to many computer vision and multimedia analysis tasks. To overcome the obstacle, some mid-level features are designed to bridge the gap, such as shape,attribute [20, 23]. These mid-level features usually need to be explicitly represented and its meaning known, which leads to another dilemma for visual problems. In recent years, the latent semantic models successfully derived in document analysis field, such as Probabilistic Latent Semantic Analysis (PLSA) [12], and Latent Dirichlet Allocation (LDA) [2], have been borrowed to address this problem. These latent semantic models can discover hidden topics/concepts in the latent semantic space based on a bag-of-words representation for the images, which can connect the low-level features and high-level semantic content. Due to the success of latent semantic models, they have been widely adopted in many applications such as image clustering, classification, and retrieval [3, 6, 18].

As we know, document analysis with PLSA and LDA models a document as a bag of words generated from a set of topics. However, it is difficult to define the visual words and their relation for image analysis. The proverb "*a picture is worth a thousand words*" shows that image is much more complex than document. Most of current applications on image analysis treat latent semantic model as a black box and each visual topic or visual word is treated independently. Not much effort has been done up to now to explore how these latent topics distribute and what correlations exist among them. In particular, there exist local structure relation among visual topics and words, which are often neglected in document analysis. Therefore, it might be unreasonable or even unfeasible to simply apply the traditional latent semantic models to visual content analysis tasks.

To address the above issues, in this paper, we present a novel probabilistic latent semantic model, named Dual Local Consistency Probabilistic Latent Semantic Analysis (DLC-PLSA), to model the latent topics with sparse neighborhood preserving embedding and local word consistency (a preliminary version appeared in [15]). Compared with the traditional PLSA models, our model has the following contributions:

- We consider the particular characteristics of image which differ from that of document, and incorporate the structure and distribution information into the traditional PLSA model.
- ℓ^1-graph is constructed to model the sparse neighborhood structure of images and embedded into topic modeling.
- the word co-occurring information is first incorporated into PLSA to help discover more accurate latent topics.

In this way, the proposed latent semantic model can estimate the probabilistic topic distributions and simultaneously consider local neighborhood structure in feature space as well as the local visual word consistency in a uniform formulation. Therefore, our model is less sensitive to noise and has more discriminative power in the latent semantic space.

The rest of the paper is organized as follows: Section 2 gives a brief review of the related work. Section 3 presents the proposed latent semantic model in detail. Extensive experimental results are reported in Section 4. Finally, we conclude our paper in Section 5.

2 Related work

Image representation is very important for image analysis and understanding tasks. Due to the semantic gap between low-level features and the semantic content of images, mid-level representations have been widely exploited by many researchers. In recent years, the latent semantic models, which can discover a latent semantic space, have attracted much attention. The latent semantic models are originally proposed for document analysis. Latent Semantic Analysis (LSA) [8], which is the first latent semantic model, uses a Singular Value Decomposition (SVD) of the term-document co-occurrence matrix to identify a latent semantic space. Despite its remarkable success in different domains, LSA has several deficits due to its unsatisfactory statistical formulation [12]. To address this issue, Hofmann proposed a generative probabilistic model named Probabilistic Latent Semantic Analysis (PLSA) [12]. PLSA models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics". However, PLSA assumes that the probability distribution of each document on the hidden topics is independent and the number of parameters in the model grows linearly with the size of corpus. Latent Dirichlet Allocation (LDA) [2] is then proposed to overcome this problem by treating the probability distribution of each document over topics as a K-parameter hidden random variable rather than a large set of individual parameters, where K is the number of hidden topics. When applying these models for image analysis, the images can be treated as documents and text words can be replaced by visual words, image regions and so on.

Probabilistic Latent Semantic Analysis (PLSA) [12] can be considered as a Bayesian network. The core of PLSA is a latent variable model for co-occurrence data which associates an unobserved topic variable $z_k \in \{z_1, \ldots, z_K\}$ with the occurrence of a word $w_j \in \{w_1, \ldots, w_M\}$ in a particular image $d_i \in \{d_1, \ldots, d_N\}$. As a generative model, PLSA is defined by the following scheme:

1. Choose an image $x_i \sim P(x_i)$,
2. Pick a latent topic $z_k \sim P(z_k|x_i)$,
3. Generate a word $w_j \sim P(w_j|z_k)$.

where $P(\cdot)$ is a probability function. As a result, one obtains an observation pair (x_i, w_j), while the latent topic variable z_k is discarded. Translating the data generation process into a joint probability model results in the expression

$$P(x_i, w_j) = P(x_i)P(w_j|x_i), \;\; P(w_j|x_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|x_i) \tag{1}$$

The parameters can be estimated by maximizing the log-likelihood

$$l = \sum_{i=1}^{N}\sum_{j=1}^{M} n(x_i, w_j) \log P(x_i, w_j) \propto \sum_{i=1}^{N}\sum_{j=1}^{M} n(x_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|x_i) \tag{2}$$

where $n(x_i, w_j)$ is the number of occurrences of term w_j in image x_i.

Latent Dirichlet allocation (LDA) [2] is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for each document w in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$,
2. Choose $\theta \sim Dir(\alpha)$,
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

where Poisson$(\cdot)$ denotes the poisson distribution, $Dir(\cdot)$ is the Dirichlet distribution, and $Multinomial(\cdot)$ is the Multinomial distribution.

However, the traditional latent semantic models treat each image or word individually in topic modeling without considering any local structure, the latent topic distributions learned by models may lack of discriminative power in many cases. In the past decade years, a variety of manifold-based approaches such as Isomap [24], Locally Linear Embedding (LLE) [22], Laplacian Eigenmaps [1] have been developed to explicitly discover the nonlinear manifold structure concealed in the data. Locality Preserving Projections (LPP) [10] and Neighborhood Preserving Embedding (NPE) [11] are the linear versions of Laplacian Eigenmaps and LLE, respectively. They are frequently added as regularization terms to many machine learning algorithms. Cai et al. propose a Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [4] model for document clustering, which models the hidden topics on document manifold by preserving the pairwise similarities of documents. Locally-consistent Topic Model (LTM) is proposed in [5], which uses KL-divergence instead of Euclidean distance to capture the document similarity. But the above methods just simply construct a K-NN graph to model the local document structure. However, as described in [7], there are some limitations of the K-NN graph. For example, it is very sensitive to data noises and one single global parameter may result in unreasonable neighborhood structure for certain data. More importantly, LapPLSI and LTM only take into account the local document structure while ignoring the word co-occurring consistency in latent semantic models. The word co-occurring consistency represents how often two words co-occur in a document or an image. If two (visual) words frequently co-occur in a document (an image), they should also to be correlated to the same topic with a high probability. Therefore, incorporating the local word consistency is also very important for modeling more accurate hidden semantic topics.

3 The proposed approach

According to recent research [1, 22, 25], image samples are often found to lie on a low-rank non-linear manifold embedded in the high-dimensional space of the original data. Therefore, exploiting the intrinsic structure concealed in the data can help discover more accurate latent topics [4, 5]. Moreover, it is reasonable to assume that visual words frequently co-occured in an image should have similar meanings or be related to similar latent topics with a high probability. Thus, the word co-occurring information is also very critical to reveal the hidden semantics in the data.

In this section, we present a novel latent semantic model, named Dual Local Consistency Probabilistic Latent Semantic Analysis (DLC-PLSA). Different from the traditional PLSA models, our model considers the sparse image neighborhood structure and local word consistency simultaneously when estimating the latent topic distributions. Therefore, it can preserve more structure semantic information in the latent semantic space.

3.1 Embedding with local structure consistency

With the PLSA model defined in Section 2, we will introduce how to embed the local structure in feature space and visual words of images into topic discovering in this subsection.

3.1.1 Sparse neighborhood consistency

In this part, we present a novel manifold learning approach based on traditional NPE method. Our method is motivated by the limitations of classical graph construction methods [22], [1] on robustness to data noise and data-adaptiveness, and recent advances in sparse coding [7], [17], [26]. With sparse representation, each sample can be reconstructed by the sparse linear superposition of the training data. The sparse reconstruction coefficients, used to deduce the weights of the ℓ^1-graph, are derived by solving an ℓ^1 optimization problem on sparse representation. Recent work in [7] has shown the ℓ^1-graph is superior to the classical graphs in various machine learning tasks such as image clustering and subspace learning.

Suppose we have an underdetermined system of linear equations: $x = D\alpha$, where $x \in R^M$ is the vector to be approximated, $\alpha \in R^N$ is the vector for unknown reconstruction coefficients, and $D \in R^{M\times N}(M < N)$ is the overcomplete dictionary with N bases. Generally, a sparse solution is more robust and is able to facilitate the consequent identification of the test sample x. In [17], L1-norm was proved to lead to a sparse solution. Like that, we seek the sparse solution to $x = D\alpha$ by solving the following optimization problem:

$$\min_{\alpha} \|\alpha\|_1, \quad s.t. \quad x = D\alpha \tag{3}$$

This problem can be solved in polynomial time by standard linear programming method. In practice, there may exist noises on certain elements of x, and a natural way to recover these elements and provide a robust estimation of α is to solve the following equation

$$x = D\alpha + \zeta = \begin{bmatrix} D & I \end{bmatrix} \begin{bmatrix} \alpha \\ \zeta \end{bmatrix} \tag{4}$$

where $\zeta \in R^M$ is the noise term. Then by setting $B = \begin{bmatrix} D & I \end{bmatrix} \in R^{M\times(M+N)}$ and $\alpha' = \begin{bmatrix} \alpha \\ \zeta \end{bmatrix}$, we can solve the following ℓ^1-norm minimization problem with respect to both reconstruction coefficients and data noises:

$$\min_{\alpha'} \|\alpha'\|_1, \quad s.t. \quad x = B\alpha' \tag{5}$$

An ℓ^1-graph summarizes the overall behavior of the whole dataset in sparse representation. The construction process is stated as follows.

1) **Inputs:** The image set denoted as the matrix $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in R^M$
2) **Robust sparse representation:** For each image x_i in the dataset, its robust sparse representation is achieved by solving the ℓ^1-norm optimization problem

$$\min_{\alpha^i} \|\alpha^i\|_1, \quad s.t. \quad x_i = B^i\alpha^i \tag{6}$$

where $B^i = [x_1, \ldots x_{i-1}, x_{i+1} \ldots, x_N, I] \in R^{M\times(M+N-1)}$ and $\alpha^i \in R^{M+N-1}$

3) **Graph weight setting:** Denote the $G = \{X, W\}$ as the ℓ^1-graph with the image set X as graph vertices and W as the graph weight matrix. We set $W_{ij} = \alpha^i_j$ if $i > j$, and $W_{ij} = \alpha^i_{j-1}$ if $i < j$.

After the ℓ^1-graph is constructed, the neighborhood structure of the image dataset as well as the graph weights is derived simultaneously in a parameter-free manner. Then, similar to NPE, sparse neighborhood preserving embedding aims to preserve the neighborhood structure of the dataset in the latent topic space by minimizing

$$R_1 = \sum_{k=1}^{K} R_{1k} = \sum_{k=1}^{K} \sum_{i=1}^{N} (P(z_k|x_i) - \sum_{j=1}^{N} W_{ij} P(z_k|x_j))^2 \tag{7}$$

where $P(z_k|x_i)$ is the probability of topic z_k given an image x_i. An intuitive explanation of minimizing R_1 is that if the image x_i can be reconstructed by its neighbors in the feature space, the intrinsic structure should also be preserved in the latent topic space.

3.1.2 Local visual word consistency

Besides the image-level local structure, the visual word consistency is usually ignored and each visual word is treated individually in the existing latent semantic models. However, the local visual word consistency is also very important for image analysis. For example, it is a natural and intuitive assumption that frequently co-occurring visual words should share similar topics in the latent space. In this part, we will introduce how to maintain the local visual word consistency in our topic model.

We first compute the co-occurrence information C_{ij} between visual word w_i and visual word w_j as follows:

$$C_{ij} = \frac{f_{ij}}{\sqrt{f_i} * \sqrt{f_j}} \tag{8}$$

where f_{ij} is the number of images in which both word w_i and word w_j appeared and f_i is the number of images in which word w_i appeared.

After we get the co-occurrence matrix C, we maintain the local word consistency in the latent topic space by minimizing

$$R_2 = \sum_{k=1}^{K} R_{2k} = \sum_{k=1}^{K} \sum_{i,j=1}^{M} (P(w_i|z_k) - P(w_j|z_k))^2 C_{ij} \tag{9}$$

An intuitive explanation of minimizing R_2 is that if the word w_i often co-occurred with w_j, their conditional distributions related to the latent topic z_k should also be similar in the latent topic space.

3.1.3 The regularized model

In order to consider the local image and visual word structure simultaneously, we add R_1 and R_2 as regularized terms to the log-likelihood of PLSA model. Then we get our new latent semantic model which aims to maximize the regularized log-likelihood as follows:

$$L = l - \lambda_1 R_1 - \lambda_2 R_2 = l - \lambda_1 \sum_{k=1}^{K} R_{1k} - \lambda_2 \sum_{k=1}^{K} R_{2k} \tag{10}$$

where l is defined by (2), in which $n(x_i, w_j)$ specifies again the number of times the visual word w_j occurred in image x_i, and $\lambda_{1,2}$ are the regularized parameters. When $\lambda_1 = \lambda_2 = 0$, our model degenerates to the traditional PLSA model. When $\lambda_1 = 0$, our model only considers the local visual word consistency. When $\lambda_2 = 0$, only the sparse neighborhood structure is preserved.

3.2 Model fitting

When a probabilistic model involves unobserved latent variables, the EM algorithm [9] is generally used for the maximum likelihood estimation of the model. EM alternates two steps: (i) an expectation (E) step where posterior probabilities are computed for the latent variables, based on the current estimates of the parameters, (ii) a maximization (M) step, where parameters are updated based on maximizing the so-called expected complete data log-likelihood which depends on the posterior probabilities computed in the E-step.

As there are regularization terms in the log-likelihood of our model, the traditional EM algorithm cannot be applied directly. Here we use the generalized EM algorithm [19] for parameter estimation. The main difference between generalized EM and traditional EM is that generalized EM algorithm finds parameters that "improve" the expected value of the log-likelihood function rather than "maximizing" it.

Let $\phi = \{P(w_j|z_k)\}$ and $\theta = \{P(z_k|x_i)\}$ denote the parameters in our model.

E-step: Our model adopts the same generative scheme as that of PLSA. Thus, we have the same E-step as that of PLSA. The posterior probabilities for latent variables are $P(z_k|x_i, w_j)$, which can be computed as follows:

$$P(z_k|x_i, w_j) = \frac{P(w_j|z_k)P(z_k|x_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|x_i)} \tag{11}$$

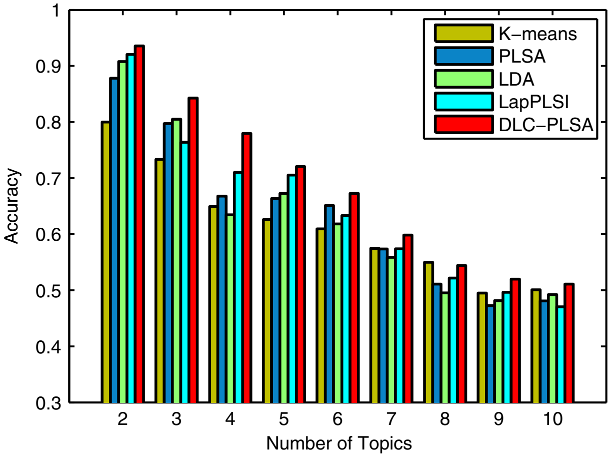


Fig. 1 Clustering results on the Binary Alphadigits

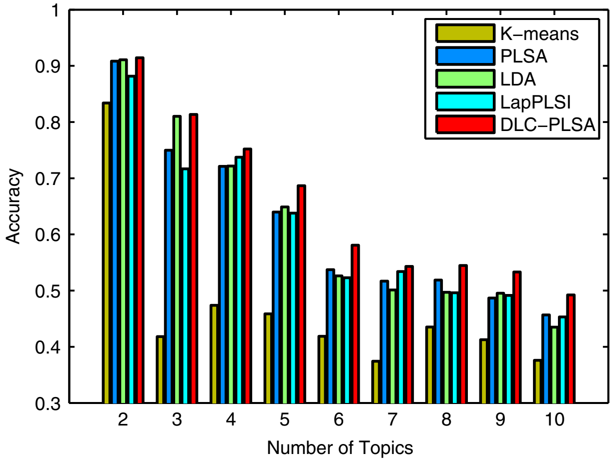


Fig. 2 Clustering results on the Scene-15 dataset

M-step: The relevant part of the expected complete log-likelihood for our model is

$$\begin{aligned} Q(\phi,\theta) &= Q_1(\phi,\theta) - \lambda_1 R_1(\theta) - \lambda_2 R_2(\phi) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(x_i,w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|x_i) \\ &- \lambda_1 \sum_{k=1}^{K}\sum_{i=1}^{N}(P(z_k|x_i) - \sum_{j=1}^{N} W_{ij}P(z_k|x_j))^2 \\ &- \lambda_2 \sum_{k=1}^{K}\sum_{i,j=1}^{M}(P(w_i|z_k) - P(w_j|z_k))^2 C_{ij} \end{aligned} \tag{12}$$

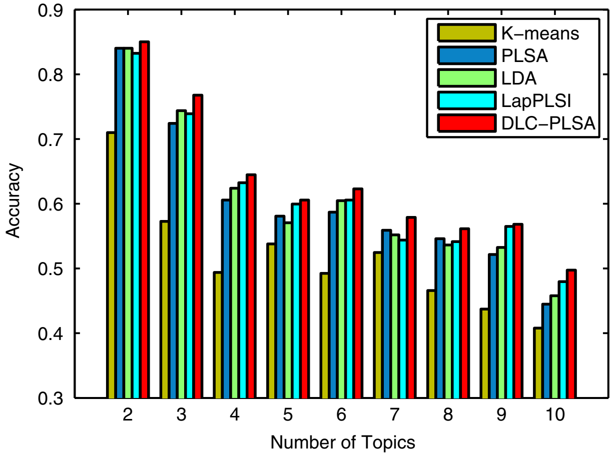


Fig. 3 Clustering results on the Caltech-101 dataset

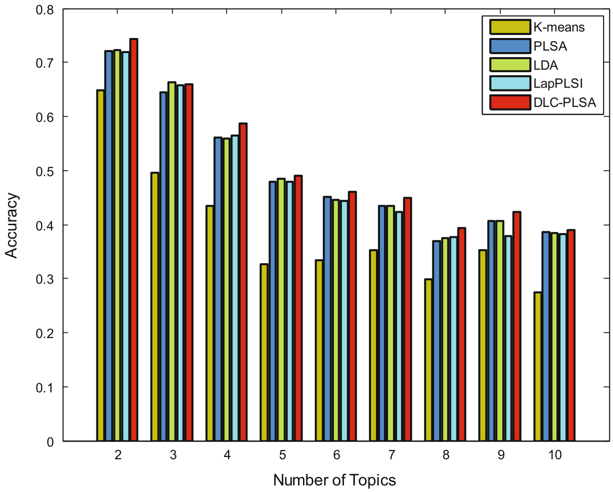


Fig. 4 Clustering results on the Caltech-256 dataset

In the M-step, we improve the expected value of the log-likelihood function $Q(\phi, \theta)$. We have parameter values $\{\phi_r, \theta_r\}$ and try to find $\{\phi_{r+1}, \theta_{r+1}\}$ which satisfy $Q(\phi_{r+1}, \theta_{r+1}) \geq Q(\phi_r, \theta_r)$ in each step.

We first find $\left\{\phi_{r+1}^{(1)}, \theta_{r+1}^{(1)}\right\}$ which maximizes $Q_1(\phi, \theta)$ instead of the whole $Q(\phi, \theta)$. This can be done by the following equations which are the M-step re-estimation of PLSA:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(x_i, w_j) P(z_k|x_i, w_j)}{\sum_{i=1}^{N} \sum_{m=1}^{M} n(x_i, w_m) P(z_k|x_i, w_m)} \tag{13}$$

$$P(z_k|x_i) = \frac{\sum_{j=1}^{M} n(x_i, w_j) P(z_k|x_i, w_j)}{\sum_{j=1}^{M} n(x_i, w_j)} \tag{14}$$

Clearly, $Q(\phi_{r+1}^{(1)}, \theta_{r+1}^{(1)}) \geq Q(\phi_r, \theta_r)$ does not necessarily hold. We then try to start from $\{\phi_{r+1}^{(1)}, \theta_{r+1}^{(1)}\}$ and decrease R_1 and R_2, which can be done through Newton-Raphson method [21]. Note that R_1 only involves parameters $P(z_k|x_i)$ while R_2 only involves parameters $P(w_j|z_k)$, we can update ϕ_{r+1} and θ_{r+1} respectively.

Given a function $f(x)$ and the initial value $x^{(t)}$, the Newton-Raphson updating formula to decrease (or increase) $f(x)$ is as follows:

$$x^{(t+1)} = x^{(t)} - \gamma \frac{f'(x^{(t)})}{f''(x^{(t)})} \tag{15}$$

Table 1 The influence of different regularized terms on the clustering accuracy of the Binary Alphadigits

Topic number	2	4	6	8
$\lambda_1 \neq 0, \lambda_2 = 0$	0.926	0.722	0.648	0.528
$\lambda_1 = 0, \lambda_2 \neq 0$	0.915	0.751	0.669	0.510
$\lambda_1 \neq 0, \lambda_2 \neq 0$	0.935	0.780	0.673	0.544

Table 2 The influence of different regularized terms on the clustering accuracy of the Scene-15 dataset

Topic number	2	4	6	8
$\lambda_1 \neq 0, \lambda_2 = 0$	0.906	0.733	0.522	0.518
$\lambda_1 = 0, \lambda_2 \neq 0$	0.907	0.679	0.553	0.503
$\lambda_1 \neq 0, \lambda_2 \neq 0$	0.914	0.752	0.581	0.545

where $0 \leq \gamma \leq 1$ is the step parameter. Since we have $R_{1k} \geq 0$, $R_{2k} \geq 0$, the Newton-Raphson method will decrease R_{1k} and R_{2k} in each updating step. With $\phi_{r+1}^{(1)}$ and put R_{1k} into the Newton-Raphson updating formula in (17), we can get the closed form solution for $\phi_{r+1}^{(2)}, \phi_{r+1}^{(3)}, \ldots, \phi_{r+1}^{(m)}$, where

$$P(z_k|x_i)_{r+1}^{(t+1)} = (1-\gamma_1)P(z_k|x_i)_{r+1}^{(t)} + \gamma_1 \sum_{j=1}^{N} W_{ij} P(z_k|x_j)_{r+1}^{(t)} \tag{16}$$

Similarly, we can also get the updating equation for θ_{r+1} as follows:

$$P(w_i|z_k)_{r+1}^{(t+1)} = (1-\gamma_2)P(w_i|z_k)_{r+1}^{(t)} + \gamma_2 \frac{\sum_{j=1}^{M} C_{ij} P(w_j|z_k)_{r+1}^{(t)}}{\sum_{j=1}^{M} C_{ij}} \tag{17}$$

Every iteration of (16) and (17) will make the topic distribution smoother. We continue the iteration of (16) and (17) until $Q\left(\phi_{r+1}^{(t+1)}, \theta_{r+1}^{(t+1)}\right) \leq Q\left(\phi_{r+1}^{(t)}, \theta_{r+1}^{(t)}\right)$. Then we test whether $Q\left(\phi_{r+1}^{(t)}, \theta_{r+1}^{(t)}\right) \geq Q(\phi_r, \theta_r)$. If not, we reject the values of $\left\{\phi_{r+1}^{(t)}, \theta_{r+1}^{(t)}\right\}$, and return the $\{\phi_r, \theta_r\}$ as the result of the M-step, and continue with the next E-step. The E-step and M-step are iteratively performed until the probability values are stable.

4 Experiments

In this section, we evaluate the performance of our model by comparing it with the state-of-the-art methods on image clustering task. Clustering is one of the most crucial techniques to organize the data samples. The latent topics discovered by latent semantic models approaches can be regarded as clusters. By representing the images in the latent space, latent semantic models can assign each image to the most probable latent topic according to the estimated conditional probability distributions $P(z_k|x_i)$. Our experiments are conducted on four publicly available datasets: the Binary Alphadigits,[1] the Scene-15 dataset [13], the Caltech-101 and Caltech-256 dataset[14]. The weighting parameters λ_1 and λ_2 are tuned with cross validation from intervals [1,100] and [1000,1500] respectively. The values of the Newton step parameter γ_1 and γ_2 are both set to 0.1 in our experiment.

The Binary Alphadigits contains binary 20x16 digits of ‘0’ through ‘9’ and capital ‘A’ through ‘Z’ where there are 39 examples of each class. Thus we have 1404 images from 36 classes in total with each image represented by a 320-dimensional binary pixel vector. Latent semantic models are applied to the images by representing each binary pixel as a word and each image as a document to generate K clusters. The Scene-15 dataset contains 4485 images from 15 categories. The Caltech-101 dataset involves 9144 images from 101

[1] http://www.cs.nyu.edu/~roweis/data.html

Table 3 The influence of different regularized terms on the clustering accuracy of the Caltech-101 dataset

Topic number	2	4	6	8
$\lambda_1 \neq 0, \lambda_2 = 0$	0.850	0.617	0.601	0.553
$\lambda_1 = 0, \lambda_2 \neq 0$	0.844	0.626	0.613	0.559
$\lambda_1 \neq 0, \lambda_2 \neq 0$	0.850	0.644	0.623	0.562

object categories and a background category, and Caltech-256 dataset consists of 256 categories and a total of 30607 images. Each image in the datasets contains a unique label to indicate which category it belongs to, which can be regarded as the ground truth for performance evaluation. SIFT [16] features are extracted and a 1000-D bag-of-words representation is generated for the Scene-15, Caltech-101 and Caltech-256 dataset. Then all the models are performed on the bag-of-words to generate K clusters. The clustering accuracy (AC) is used to measure the clustering performance [27].

We evaluated the proposed DLC-PLSA model and compared it with the following related algorithms:

- K-means clustering algorithm (K-means),
- Probabilistic Latent Semantic Analysis (PLSA) [12],
- Latent Dirichlet Allocation (LDA)[2],
- Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [4].

In order to make the experiments statistically meaningful, we conducted the evaluations with the cluster numbers ranging from two to ten. For each given number of clusters k, we implement five test runs. In each test run k different categories were randomly selected from the datasets and provided to the clustering algorithms. The final performance scores for each k were obtained by averaging the scores over its five test runs.

Figures 1, 2, 3 and 4 shows the clustering performance of all the algorithms on the Binary Alphadigits, the Scene-15 dataset, the Caltech-101 dataset and Caltech-256 dataset, respectively. We can see that the latent models PLSA and LDA achieve better performance than the traditional K-means clustering method on the whole. But the PLSA and LDA model show lower performance than LapPLSI and DLC-PLSA because they do not consider any local discriminant structure when discovering the latent topics. Although the LapPLSI model considers the proximity between image pairs, it is not robust enough and sometimes even gets worse results than PLSA and LDA. The reason is that a K-NN graph is simply constructed to model the image structure, which is very sensitive to noise, and the local word consistency is also ignored. Therefore, it cannot reach full discriminant power. Our DLC-PLSA model, which constructs the ℓ^1-graph to model the neighborhood structure of images and incorporates the local word consistency into the model at the same time, can perform consistently better than other models.

Table 4 The influence of different regularized terms on the clustering accuracy of the Caltech-256 dataset

Topic number	2	4	6	8
$\lambda_1 \neq 0, \lambda_2 = 0$	0.7178	0.5609	0.4456	0.37
$\lambda_1 = 0, \lambda_2 \neq 0$	0.714	0.5471	0.4517	0.3618
$\lambda_1 \neq 0, \lambda_2 \neq 0$	0.7432	0.5876	0.4597	0.3928

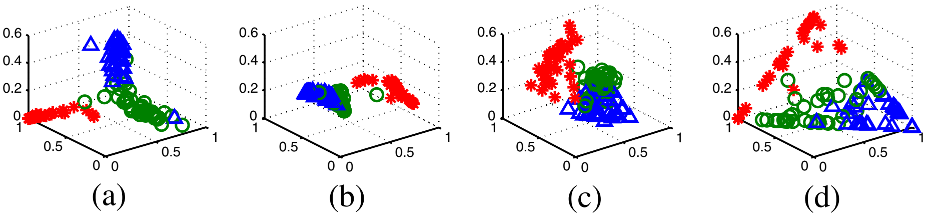


Fig. 5 Visualization view of image distribution in the latent topic space learned by different topic models. **a** The DLC-PLSA model. **b** The LapPLSI model. **c** The LDA model. **d** The PLSA model. (digit characters ‘A’ - ‘C’ in the Binary Alphadigits dataset, ‘A’: *blue*, ‘B’: *green*, ‘C’: *red*)

In order to evaluate the performance of different regularized terms, we also compare the results of our model with different regularized terms by setting the regularization parameter $\lambda_1 = 0$ or $\lambda_2 = 0$ respectively on different topic numbers. The comparison results are shown in Tables 1, 2, 3 and 4, from which we can see that the model including both regularized terms always performs as well as or outperforms the best one with only one regularized term. Moreover, the model with only the visual word regularized term also performs very consistently and sometimes gets better results than the image regularized term, which proves that the local visual word consistency is also very important in latent semantic models.

The visualization comparison of image distribution (digit characters ‘A’ - ‘C’ in the Binary Alphadigits dataset) in the latent topic space is shown in Fig. 5. The comparison results show that the embedded representations of DLC-PLSA, which models the hidden topics with sparse neighborhood and local word consistency, have the best separability. Although LapPLSI considers the proximity of image pairs in topic modeling, it cannot separate characters ‘A’ and ‘B’ very well. We analyzed the reason and found that three ‘noisy’ images of character ‘B’ were very easily considered as neighbors by most images of character ‘A’ when constructing the K-NN graph, which affects the whole local structure significantly.

In order to further show the impact of the noisy images on the performance of different models, we conduct an experiment by manually removing some noisy images in these three classes and the comparison clustering results are given in Table 5. We can see that the performance of LapPLSI is affected greatly by the noisy images, because it models the image structure on the K-NN graph. After removing the noisy images, its performance has a big improvement. In contrast, our model constructs ℓ^1-graph to model the neighborhood structure of images and it can perform very consistently in both cases, which indicates that our model is more robust to noise and can discover more accurate latent topics.

Table 5 The impact of noisy images on clustering accuracy (digit characters ‘A’, ‘B’, ‘C’) of different models

	DLC-PLSA	LapPLSI	LDA	PLSA
Without removing the noisy images	0.8718	0.6667	0.7949	0.7179
Removing the noisy images	0.8803	0.8034	0.8291	0.8205

5 Conclusion

In this paper, we have presented a novel latent semantic model named DLC-PLSA. In our DLC-PLSA, the sparse neighborhood structure of the images and the local word consistency are preserved when modeling the latent topics. Therefore, our model can be less sensitive to noise and have more powerful description ability in the latent semantic space than the traditional latent semantic models. Experimental results on image clustering show that the DLC-PLSA model provides better representation and gets very consistent performance.

There are also some questions remained to be investigated in the future work. First, the idea of dual local consistency preserving can also be incorporated into other clustering methods or topic models. Second, the ℓ^1-graph is unsupervised in our model. The label information can be used to construct a more discriminative graph. Finally, it is very interesting to explore new ways in order to capture the image or word correlations effectively.

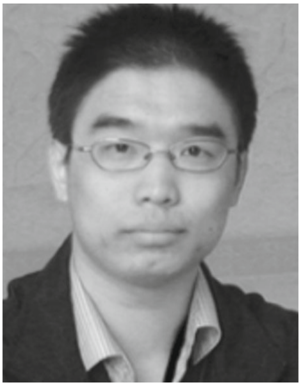
Acknowledgments This work was supported in part by 973 Program (Grant No. 2010CB327905), National Natural Science Foundation of China (Grant No. 61332016, 61202325).

References

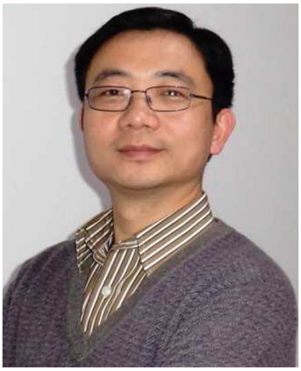
1. Belkin M, Niyogi P (2002) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373–1396
2. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
3. Bosch A, Zisserman A, Munoz X (2006) Scene classification via PLSA. In: Proceedings of ECCV. pp 517–530
4. Cai D, Mei Q, Han J, Zhai C (2008) Modeling hidden topics on document manifold. In: Proceedings of CIKM. pp 911–920
5. Cai D, Wang X, He X (2009) Probabilistic dyadic data analysis with local and global consistency. In: Proceedings of ICML
6. Cao L, Li F (2007) Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: Proceedings of ICCV. pp 1–8
7. Cheng B, Yang J, Yan S, Fu Y, Huang T (2010) Learning with .1-graph for image analysis. IEEE Trans Image Process 19:858–866
8. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407
9. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from in complete data via the EM algorithm. J Royal Stat Soc 39:1–38
10. He X, Niyogi P (2003) Locality preserving projections. In: Proceedings of NIPS
11. He X, Cai D, Yan S, Zhang H (2005) Neighborhood preserving embedding. In: Proceedings of ICCV. pp 1208–1213
12. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42:177–196
13. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR. pp 2169–2178
14. Li F, Rob F, Pietro P (2004) Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: Proceedings of CVPR workshop on generative model based vision
15. Li P, Cheng J, Lu H (2012) Modeling hidden topics with dual local consistency for image analysis. In: Proceedings of ACCV. pp 648–659
16. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
17. Meinshansen N, Buhlmann P (2006) High-dimensional graphs and variable selection with the lasso. Annals Stat 34:1436–1462
18. Monay F, Gatica-Perez D (2004) PLSA-based image auto-annotation: constraining the latent space. In: Proceedings of ACM multimedia. pp 348–351

19. Neal R, Hinton G (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. Learn Graph Models
20. Parikh D, Grauman K (2011) Relative attributes. In: Proceedings of IEEE international conference on computer vision
21. Press W, Flannery B, Teukolsky S, Vetterling W (1992) Numerical recipes in C: the art of scientific computing. Cambridge University Press
22. Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
23. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
24. Tenenbaum J (1997) Mapping a manifold of perceptual observations. In: Proceedings of NIPS. pp 682–688
25. Tenenbaum J, Silva V, Langford J (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
26. Wright J, Genesh A, Yang A, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31:210–227
27. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of SIGIR. pp 267–273

Jian Cheng is currently an associate professor of Institute of Automation, Chinese Academy of Sciences. He received the B.S. and M.S. degrees in Mathematics from Wuhan University in 1998 and in 2001, respectively. In 2004, he got his Ph.D degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences. From 2004 to 2006, he has been working as postdoctoral in Nokia Research Center. Then he joined National Laboratory of Pattern Recognition, Institute of Automation. His current research interests include image and video search, machine learning, etc. He has authored or co-authored more than 50 academic papers and co-edited two books in these areas. He was the recipient of LU JIAXi Young Talent award in 2010. Dr. Cheng served as Program co-chair for ACM International Conference on Internet Multimedia Computing and Services (ICIMCS'10), Technical Program Committee member for ACM Multimedia 2008-2009, IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), IEEE International Conference on Multimedia and Expo (ICME'08-12), IEEE International Conference on Computer Vision (ICCV'07), etc. He has also co-organized one special issue on Pattern Recognition Journal.



Peng Li received the B.S. degree in automation from Shandong University, Jinan, China, in 2008. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia retrieval, image and video analysis, and machine learning.



Ting Rui received the M.S. degree and Ph.D. from PLA University of Science and Technology, Nanjing, China, in 1998 and 2001respectively. Ting RUI is Professor of Information Technology Department in the PLA University of Science and Technology. He mainly applies computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored more than 80 scientific articles.

Hanqing Lu received the B.E. and M.E. degrees from Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China in 1992. Currently, he is a Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include computer vision, object tracking, recognition and image retrieval, etc. He has published more than 200 papers in those areas.