An Extended Generative Feature Learning Algorithm for Image Recognition

Bin Wang[1], Chuanjiang Li[1*], Qian Zhang[1], Jifeng Huang[1]
[1] College of Information, Mechanical and Electrical Engineering, Shanghai Normal University,
Shanghai, 200234, China
[e-mail : lichuanjiang2016@hotmail.com]
*Corresponding author: Chuanjiang Li

Received March 31, 2017; revised May 25, 2017; accepted June 16, 2017;
published August 31, 2017

Abstract

Image recognition has become an increasingly important topic for its wide application. It is highly challenging when facing to large-scale database with large variance. The recognition systems rely on a key component, i.e. the low-level feature or the learned mid-level feature. The recognition performance can be potentially improved if the data distribution information is exploited using a more sophisticated way, which usually a function over hidden variable, model parameter and observed data. These methods are called generative score space. In this paper, we propose a discriminative extension for the existing generative score space methods, which exploits class label when deriving score functions for image recognition task. Specifically, we first extend the regular generative models to class conditional models over both observed variable and class label. Then, we derive the mid-level feature mapping from the extended models. At last, the derived feature mapping is embedded into a discriminative classifier for image recognition. The advantages of our proposed approach are two folds. First, the resulted methods take simple and intuitive forms which are weighted versions of existing methods, benefitting from the Bayesian inference of class label. Second, the probabilistic generative modeling allows us to exploit hidden information and is well adapt to data distribution. To validate the effectiveness of the proposed method, we cooperate our discriminative extension with three generative models for image recognition task. The experimental results validate the effectiveness of our proposed approach.

Keywords: Feature learning , distribution modeling, generative score space

https://doi.org/10.3837/tiis.2017.08.013 **ISSN : 1976-7277**

1. Introduction

Image recognition has been a popular research subject in computer vision with applications in image retrieval systems, vehicle navigation, and video analysis [1]. Among a few possible taxonomies, the literature present two orthogonal methods: the discriminative and the generative paradigms. Specifically, the discriminative approaches present the boundaries between different classes, instead of modeling the distribution of samples belonging to the same class. These methods aim to maximize the data separately. Discriminative models directly model the map from images to classes, by capturing the decision bounds among different classes. Specifically, these approaches learn a separate classifier for each class. The classifier is then used to predict whether the class can be assigned to the test image [2][3]. A variety of discriminative models, such as support vector machines (SVM) [4], discriminative kernel type model [5], and multiple-instance learning [6] have been applied to image recognition. On the other hand, the generative approaches model the distribution of data and tell people prior knowledge by a graph structure, which establishes correspondences between image feature and the model by means of conditional distribution. The generative model integrates hidden variables and is good at dealing with missing data, especially when little labeled data is available. Generative models model the distribution of images and explain how they can be generated. They can infer and exploit information hidden in images. The hidden information usually closely relates to the high level concepts of images. For instance, in probabilistic latent semantic analysis (pLSA) [7], the latent space representation can capture high-level relations within and across the class and visual modalities. Generative models therefore can utilize this additional high level information for image recognition. By means of naive Bayes classifier, they can be used to perform classification.

Different from the above approaches, a probabilistic branch of methods, generative score spaces[8][9], recently received increasing attention in image recognition. These methods derive the explicit feature mapping based on the probabilistic distribution over the data. Consequently, they are able to exploit the abilities of probability models, e.g. dealing with structured data and exploiting hidden variables. And then the derived features are straightforwardly delivered to discriminative models for image classification. So, these approaches can simultaneously benefit from the advantages of the two paradigms. The representative methods include probability product kernels [10], Kullback Leibler divergence based similarity [11], Fisher kernel [12], free energy score space [8], and posterior divergence [9]. However, all the above methods derive feature mappings from the generative model trained using samples from all classes, without making use of the class label.

In this paper, we propose a discriminative extension of the existing generative score space methods for the image recognition task. The proposed approach can exploit class label when deriving score function (i.e. feature mapping), which is validated very informative in image classification. Specifically, we first extend the regular generative models to class conditional models over both observed variable and class label. Then, we derive the feature mapping from the extended generative models. And at last the derived features are straightforwardly delivered to discriminative models for image classification. The proposed extension models the joint distribution of the observed data $\mathbf{x}$ and its label y, and applies three representative score space methods, i.e. free energy score space (FESS) [8] posterior divergence (PD) [9], and sufficient statistics (SS) [13], to the joint model respectively. In a word, different from the previous approaches, there are three advantages of our proposed approach: first, the resulted feature mappings are very simple, i.e., the concatenation of the weighted feature mappings respectively derived from the class-conditional models; second, the probabilistic generative modeling allows us to exploit hidden information and is well adapt to data distribution; third, the joint model fully exploits class label information in an effective way and the derived features are fixed length.

The remainder of this paper is organized as follows. Section 2 presents the preliminaries and revisits some representative generative score space methods. Section 3 presents the framework of our proposed discriminative extension for existing generative score spaces. Section 4 experimentally evaluates the model for image recognition. Section 5 draws a conclusion.

2. Preliminaries and Generative Score Space Revisit

2.1 Generative Score Space

Generative score space is a branch of probabilistic methods, which derive the explicit feature mapping based on the probabilistic distribution over the data from generative score space [14]. Generative score space methods are first proposed in [15]. These methods can be categorized into two classes [16] [17]: parameter-based methods and random variable based methods. Specifically, let $P(\mathbf{x}|\theta)$ be the marginal distribution of an adopted generative model, where $\mathbf{x}\in\mathbf{R}^D$ is the observed variable and $\theta=\{\theta_1,\cdots,\theta_K\}$ is the set of K parameters. Parameter based methods are represented by Fisher Score (FS) [12], which derives explicit feature mappings from a given generative model. The feature mappings measure how a sample affects the parameter θ, i.e., differential operation over the log likelihood $\log P(\mathbf{x}|\theta)$ with respect to parameters. The advantage of this method is that it is robust to the number of hidden variables [16]. Also, it shows state-of-the-art performance in image recognition [18][19]. Variable based methods are represented by free energy score space (FESS) [8], posterior divergence (PD) [9] and augmented sufficient statistics (SS) [13]. For example, PD derives feature mappings by mainly measuring how well a sample fits the distribution of the random variables.

Generative score space methods work on the variational lower bound of the log likelihood function of generative models. Suppose $\mathbf{x}\in\mathbf{R}^D$ be the observed variable and $\{\mathbf{x}^i\}_{i=1}^N$ be N training samples. Let the probabilistic distribution over x is modeled by a generative model with hidden variables H introduced, and the model is parameterized by a vector of parameters $\boldsymbol{\theta}$. Suppose $P(\mathbf{x},H|\boldsymbol{\theta})$ be the joint distribution and $P(\mathbf{x}|\boldsymbol{\theta})$ be the marginal distribution. The marginal distribution $P(\mathbf{x}|\boldsymbol{\theta})$ is unavailable for most cases since the integration $\int P(\mathbf{x},H|\boldsymbol{\theta})dH$ is intractable [20]. The common idea of methods which are developed to attack this problem is to construct an approximate posterior distribution $Q^t(H)$ to estimate the real posterior distribution. Then we have :

$$\begin{aligned}&\log P(\mathbf{x}^t|\theta)\\&=-\mathbf{KL}(Q^t(H)\|P(\mathbf{x}^t,H|\theta))+\mathbf{KL}(Q^t(H)\|P(H|\mathbf{x}^t,\theta))\end{aligned} \tag{1}$$

where **KL** denotes the Kullback–Leibler divergence, and the second term measures the residual error of using $Q^t(H)$ to approximate $P(H|\mathbf{x}^t,\theta)$; and takes zero when $Q^t(H)$ is expressive enough. In this case, the first term is the exact log likelihood. In this paper, we focus on the variational inference [20] which resorts to a lower bound of the log likelihood as follows:

$$\log P(\mathbf{x}^t|\theta) \geq -\mathbf{KL}(Q^t(H) \| P(\mathbf{x}^t, H|\theta)) = -F(Q^t, \theta) \tag{2}$$

$$\begin{aligned} F(Q,\theta) &= \sum_i F(Q^i, \theta) \\ &= \sum_i E_{Q^i}[\log Q^i(H) - \log P(\mathbf{x}, H|\theta)] \end{aligned} \tag{3}$$

where $-F(Q^t,\theta)$ is the variational lower bound; $F(Q^t,\theta)$ is the free energy function. A choice for $Q^t(H)$ is that it takes the same form with $P(H)$ but with different parameters [20]. The above formulation involves two approximations which are using the approximate posterior distribution $Q^t(H)$ to approach the real posterior distribution $P(H|\mathbf{x}^t,\theta)$ and using the lower bound $-F(Q^t,\theta)$ to approach the real log likelihood $\log P(\mathbf{x}^t|\theta)$. The above two approximations will not lose generality. That is because when $Q^t(H)$ is given by exact inference methods, the approximate posterior $Q^t(H)$ exactly equals to the real posterior $P(H|\mathbf{x}^t,\theta)$, and the lower bound $-F(Q^t,\theta)$ exactly equals to the real log likelihood $\log P(\mathbf{x}^t|\theta)$. In the following section, we will review three representative score space approaches: free energy score space (FESS) [8], posterior divergence (PD) [9] and sufficient statistics (SS) [13].

2.2 Free Energy Score Space (FESS)

Suppose $H=\{h_1,\cdots,h_M\}$ be a set of M hidden variables of the adopted generative model. Let pa_x be the parent variables of $\mathbf{x}$, and $P(\mathbf{x}|pa_x)$ model the relationship between $\mathbf{x}$ and pa_x. Let pa_i be the parent variables of h_i, and $P(h_i|pa_x)$ model the relationship between h_i and pa_i. pa_x and pa_i can be null. For directed graphical models, we have the following expression:

$$\begin{cases} P(\mathbf{x}, H|\theta) = P(\mathbf{x}|pa_x)\prod_i P(h_i|pa_i) \\ \quad Q^t(H) = \prod_i Q^t(h_i|pa_i) \end{cases} \tag{4}$$

Substituting the above factorization into the free energy function $F(Q^t,\theta)$, free energy score space (FESS) [8] expands the free energy function according to the summation terms, and uses the resulting terms as the score functions.

$$
\begin{aligned}
&F(Q^t,\theta)\\
&=\mathrm{E}_{Q^t(H)}[\log Q^t(H)-\log P(\mathbf{x}^t,H\,|\theta)]\\
&=\mathrm{E}_{Q^t(H)}\left[\log\prod_i Q^t(h_i\,|pa_{h_i})-\log P(\mathbf{x}^t\,|pa_x)\prod_i P(h_i\,|pa_i)\right]\\
&=\mathrm{E}_{Q^t(H)}[\log Q^t(h_1\,|pa_{h_i})]+\cdots+\mathrm{E}_{Q^t(H)}[\log Q^t(h_M\,|pa_{h_M})]\\
&\quad-\mathrm{E}_{Q^t(H)}[\log P(\mathbf{x}^t\,|pa_x)]\\
&\quad-\mathrm{E}_{Q^t(H)}[\log P(h_1\,|pa_1)]-\cdots-\mathrm{E}_{Q^t(H)}[\log P(h_M)]\\
&=\sum_i f_i(\mathbf{x}^t)=\alpha_{\mathrm{FESS}}^T\Phi_{\mathrm{FESS}}(\mathbf{x}^t)
\end{aligned}
\tag{5}
$$

where $\alpha_{\mathrm{FESS}}=(1,\cdots,1)^T$; $f_i(\mathbf{x}^t)$ is the i-th element of the score function of FESS; the parent variable of h_M is null. The complete score function of FESS can be written as:

$$
\Phi_{\mathrm{FESS}}(\mathbf{x}^t)=(f_1(\mathbf{x}^t),\cdots,f_M(\mathbf{x}^t))^T \tag{6}
$$

2.3 Posterior Divergence (PD)

Suppose $\mathbf{x}^t$ be the observed sample at the n-th iteration; χ=($\mathbf{x}^1,\cdots,\mathbf{x}^{N-1}$) be a set of samples not containing $\mathbf{x}^t$; $\chi_{+t}=\chi\cup\{\mathbf{x}^t\}$ be the resulting set by adding $\mathbf{x}^t$ to χ; $P(\mathbf{x}|\theta)$ be the model estimated from the set χ and $Q^i(H)$ be the approximations of posterior distribution $P(H\,|\mathbf{x}^i,\theta)$ where $\mathbf{x}^i\in\chi$; $P(\mathbf{x}|\theta_{+t})$ be the model estimated from χ_{+t}; $Q^i_{+t}(H)$ be the approximations of posterior distribution $P(H\,|\mathbf{x}^i,\theta_{+t})$ where $\mathbf{x}^i\in\chi_{+t}$. The implied log likelihood of $\mathbf{x}^t$ derived in incremental EM algorithm could be written as the contribution of $\mathbf{x}^t$ to the log likelihood for the whole sample set:

$$
\begin{aligned}
&L(\mathbf{x}^t)\\
&=\sum_{i=1}^{N}[-F(Q^i_{+t},\theta_{+t})]-\sum_{i=t}^{N}[-F(Q^i,\theta)]
\end{aligned}
\tag{7}
$$

This log likelihood is different from the previous log likelihood, i.e. lower bound $-F(Q^t,\theta)$ in variational EM algorithm. Substituting Eq. (3) into Eq. (7), the expansion of the implied log likelihood can be obtained. To derive score function, the factorization can be obtained from the resulting expansion, as shown in Eq. (4). Substitute Eq. (4) into Eq. (7), the score function can be obtained as follows:

$$
\begin{aligned}
L(\mathbf{x}^t)\approx{}&[\underbrace{\sum_{i\neq t}^{N}E_{Q^i}\log\frac{P(\mathbf{x}^i\,|pa_x,\theta_{+t})}{P(\mathbf{x}^i\,|pa_x,\theta)}}_{\varphi_{pd}^{\mathbf{x}}}+\underbrace{E_{Q^t}\log P(\mathbf{x}^t\,|pa_x,\theta_{+t})]}_{\varphi_{fit}^{\mathbf{x}}}\\
&+[\underbrace{\sum_{i\neq c}^{N}E_{Q^i}\log\frac{P(h_1\,|pa_1,\theta_{+t})}{P(h_1\,|pa_1,\theta)}}_{\varphi_{pd}^{h_1}}+\underbrace{E_{Q^t}\log P(h_1\,|pa_1,\theta_{+t})]}_{\varphi_{fit}^{h_1}}\\
&-\underbrace{E_{Q^t}\log Q^t(h_1\,|pa_1)}_{\varphi_{ent}^{h_1}}+\underbrace{\cdots}_{h_2\cdots h_{M-1}}+[\underbrace{\cdots}_{\varphi_{pd}^{h_M}}+\underbrace{\cdots}_{\varphi_{fit}^{h_M}}-\underbrace{\cdots}_{\varphi_{ent}^{h_M}}]
\end{aligned}
\tag{8}
$$

where pa_x is the parent variables of $\mathbf{x}$, pa_m is the parent variables of hidden variable h_m, $m = 1, \cdots, M$. For the input sample $\mathbf{x}^t$, the complete score function of PD is:

$$\Phi_{PD}(\mathbf{x}^t) = (\varphi_{pd}^{\mathbf{x}}, \varphi_{fit}^{\mathbf{x}}; \varphi_{pd}^{h_1}, \varphi_{fit}^{h_1}, \varphi_{ent}^{h_1}; \cdots; \varphi_{pd}^{h_M}, \varphi_{fit}^{h_M}, \varphi_{ent}^{h_M}) \tag{9}$$

2.4 Sufficient Statistics (SS)

Suppose $\mathbf{x}$ be the observed variable. The joint distribution $P(\mathbf{x}, H|\boldsymbol{\theta})$ of an adopted generative model can be expressed as follows [13]:

$$P(\mathbf{x}, H|\boldsymbol{\theta}) = \exp\{\alpha(\boldsymbol{\theta})^T T(\mathbf{x}, H) + A(\boldsymbol{\theta})\} \tag{10}$$

where $\alpha(\boldsymbol{\theta})$ is a function vector defined on model parameter $\boldsymbol{\theta}$; $T(\mathbf{x}, H)$ is a vector constructed by sufficient statistics functions defined on $\mathbf{x}$ and H; $A(\boldsymbol{\theta})$ is a scalar function of defined on $\boldsymbol{\theta}$. Considering $P(\mathbf{x}, H) = P(\mathbf{x}|H)P(H)$, we have:

$$P(H|\boldsymbol{\theta}_h) = \exp\{\alpha(\boldsymbol{\theta}_h)^T T(H) + A(\boldsymbol{\theta}_h)\} \tag{11}$$

where $P(H)$ obeys exponential family distribution the same as the generative model.

For a sample $\mathbf{x}^t$, the approximate posterior distribution $Q^t(H)$ have the same form with prior distribution $P(H|\boldsymbol{\theta}_h)$. That is

$$Q^t(H) = \exp\{\alpha(\boldsymbol{\theta}_h^t)^T T(H^t) + A(\boldsymbol{\theta}_h^t)\} \tag{12}$$

where $\boldsymbol{\theta}_h^t$ is the parameter vector depending on the sample $\mathbf{x}^t$. Substituting Eq. (10) and Eq. (12) into Eq. (2), we have

$$\begin{aligned}
&F^t(Q, \theta) \\
&= \mathrm{E}_{Q^t(H)}[\alpha(\boldsymbol{\theta})^T T(\mathbf{x}^t, H^t) + A(\boldsymbol{\theta}) - \alpha(\boldsymbol{\theta}_h^t)^T T(H^t) - A(\boldsymbol{\theta}_h^t)] \\
&= \mathrm{E}_{Q^t(H)}[\alpha(\boldsymbol{\theta})^T T(\mathbf{x}^t, H^t) - \mathbf{1}^T diag(\alpha(\boldsymbol{\theta}_h^t))T(H^t) - A(\boldsymbol{\theta}_h^t) + A(\boldsymbol{\theta})] \\
&= \alpha(\boldsymbol{\theta})^T \mathrm{E}_{Q^t(H)}[T(\mathbf{x}^t, H^t)] - \mathbf{1}^T diag(\alpha(\boldsymbol{\theta}_h^t))\mathrm{E}_{Q(H^t)}[T(H^t)] \\
&- A(\boldsymbol{\theta}_h^t) + A(\boldsymbol{\theta})] \\
&= \eta^T \mathrm{E}_{Q^t(H)}[\varphi(\mathbf{x}^t, H^t)] = \eta^T \Phi(\mathbf{x}^t)
\end{aligned} \tag{13}$$

where $\eta = (\alpha(\boldsymbol{\theta})^T, -\mathbf{1}^T, -1, A(\boldsymbol{\theta})^T$ only depends on parameter $\boldsymbol{\theta}$; $\varphi^t(\mathbf{x}^t, H^t)$ is a function over observe variable $\mathbf{x}^t$ and hidden variable H^t:

$$\varphi(\mathbf{x}^t, H^t) = (T(\mathbf{x}^t, H^t)^T, (diag(\alpha(\boldsymbol{\theta}_h^t))T(H^t))^T, A(\boldsymbol{\theta}_h^t), 1)^T \tag{14}$$

For an input sample $\mathbf{x}^t$, the complete score function of SS is:

$$\Phi_{SS}(\mathbf{x}^t) = E_{Q(H^t)}[\varphi(\mathbf{x}^t, H^t)] \tag{15}$$

Although the above three approaches can exploit the information of hidden variables and data distribution, they do not exploit the class label, which is demonstrated very informative in recognition. To overcome the limitations of previous generative score space methods, we propose a new method in the next section.

3. Feature Learning From Generative Score Space Extension

3.1 Model Formulation

Let the input sample be $\mathbf{x} \in \mathrm{R}^D$ and its label be $y \in \{1, \cdots, C\}$. We model the joint distribution of $\mathbf{x}$ and y. It is worth noting that the labels of training samples are known and labels for test samples are unknown. So we can treat y as a random variable, which follows the Multinomial distribution. We have:

$$P(y) = \prod_{c=1}^{C} \alpha_c^{I(y=c)} \tag{16}$$

where $I(y=c)$ is an indication function, which outputs 0 if $y=c$ is false and outputs 1 if $y=c$ is true. α_c is the mixing prior satisfying $\alpha_c = E_{P(y)}[I(y=c)]$. The distribution over $\mathbf{x}$ and H, conditioned on y, is as follows:

$$P(\mathbf{x}, H \left| y \right.) = \prod_{c=1}^{C} P(\mathbf{x}, H \left| y=c \right.)^{I(y=c)} = \prod_{c=1}^{C} P(\mathbf{x}, H \left| \theta_c \right.)^{I(y=c)} \tag{17}$$

where C is the number of class label. Here, as mentioned in the above section, we assume that $P(\mathbf{x}, H \left| \theta_c \right.)$ is the joint distribution of a given generative model with hidden variable H and model parameter θ_c.

The joint distribution over $\mathbf{x}$, H and y is the multiplication of Eq. (16) and Eq. (17):

$$\begin{aligned} P(\mathbf{x}, H, y) &= P(\mathbf{x}, H \left| y \right.) \cdot P(y) \\ &= \prod_{c=1}^{C} P(\mathbf{x}, H \left| \theta_c \right.)^{I(y=c)} \cdot \prod_{c=1}^{C} \alpha_c^{I(y=c)} \end{aligned} \tag{18}$$

As mentioned above, we treat y as a hidden variable. If y is assigned an exact value, for example, $y = y_0$ for $\mathbf{x}^t$, indicates that $P(y = y_0 \left| \mathbf{x}^t \right.) = 1, P(y \neq y_0 \left| \mathbf{x}^t \right.) = 0$, $Q^t(y = y_0) = 1$ and $Q^t(y \neq y_0) = 0$. If y is unknown, we can infer its posterior using methods suggested by [20] as follows:

$$Q^t(H) = \prod_{c=1}^{C} Q_c^t(H)^{I(y=c)} \text{ and } Q^t(y) = \prod_{c=1}^{C} \gamma_c^{I(y=c)} \tag{19}$$

where we choose the approximate posterior $Q^t(H)$ and $Q^t(y)$ as the same form of $P(H)$ and $P(y)$ respectively, and $\gamma_c^t = \mathrm{E}_{Q^t(y)}[I(y=c)]$ is the expectation of class label. The log likelihood of the above model can be written as:

$$
\begin{aligned}
\log P(\mathbf{x}^t) &\geq -\mathrm{KL}(Q^t(H)Q^t(y)\|P(\mathbf{x}^t, H, y)) \\
&= -\mathrm{KL}(Q^t(y)\prod_c Q_c^t(H)^{I(y=c)} \| P(y)\prod_c P(\mathbf{x}^t, H|\theta_c)^{I(y=c)}) \\
&= \sum_c -\mathrm{KL}(Q_c^t(H)\|P(\mathbf{x}^t, H|\theta_c))\cdot\gamma_c^t - \mathrm{KL}(\mathrm{Q}^t(\mathrm{y})\|P(y)) \\
&= \sum_c -\alpha_c^T \Phi_c(\mathbf{x}^t)\cdot\gamma_c^t - \mathrm{KL}(\mathrm{Q}^t(\mathrm{y})\|P(y)) \\
&= \alpha^T \Phi(\mathbf{x}^t)
\end{aligned}
\tag{20}
$$

where $\alpha_c^T \Phi_c(\mathbf{x}^t)$ can be any expression of FESS in Eq.(5) or the expression of PD in Eq. (8) or the expression of SS in Eq. (15) over the c-th class-conditional model; $\alpha = (-\alpha_1^T, \ldots, -\alpha_c^T, -1)^T$. The derived score function is as follows:

$$
\Phi(\mathbf{x}^t) = (\gamma_1^t \Phi_1(\mathbf{x}^t)^T, \cdots, \gamma_c^t \Phi_c(\mathbf{x}^t)^T, \mathrm{KL}(\mathrm{Q}^t(\mathrm{y})\|P(y)))^T \tag{21}
$$

Because $\mathrm{KL}(\mathrm{Q}^t(\mathrm{y})\|P(y))$ is less informative, the derived score function can be roughly considered as the weighted version of previous score function over C models. A procedure implied in our approach is to infer the label for an input sample $\mathbf{x}^t$. This is realized by the estimation of the posterior distribution over class label:

$$
Q^t(y=c) = \gamma_c^t = \frac{\alpha_c P(\mathbf{x}^t|\theta_c)}{\sum_i \alpha_i P(\mathbf{x}^t|\theta_i)} \tag{22}
$$

γ_c^t takes a large value when the generative model believes the input $\mathbf{x}^t$ is likely to have label c and gives the score function over θ_c a larger weight. Our approach takes both the class label information and the model's guess into account by means of such weights γ_c^t.

The training procedure of our methods is summarized in **Algorithm 1**, and the derivation of score function is summarized in **Algorithm 2**.

Algorithm 1. Train the joint model

```
input: training set $\{(\mathbf{x}^t, y^t)\}_{t=1}^N$ of $C$ classes
initialize $\{\theta_c\}_{c=1}^C$ and $\alpha=(\alpha_1,\cdots,\alpha_C)^T$
for $i=1$ to $C$ do
    training model $\theta_c$ using the samples of the class $c$
    $\alpha_c = \frac{1}{N}\sum_{t=1}^N I(y^t=c)$
end for
output: $\{\theta_c\}_{c=1}^C$ and $\alpha$
```

Algorithm 2. Compute score functions.

```
input: parameters $\{\theta_c\}_{c=1}^{C}$, $\alpha$ and dataset $\{\mathbf{x}^t\}_{t=1}^{N_f}$
for $t = 1$ to $N_f$ do
    $Q_c^t(H) = \max_{Q_c^t(H)} KL(Q_c^t(H) \| P(\mathbf{x}^t, H | \theta_c))$
    compute score space $\Phi_c^t(\mathbf{x}^t)$ using FESS, SS, PD
    $\gamma_c^t = \mathrm{E}_{Q^t(y)}[I(y = c)] = \alpha_c P(\mathbf{x}^t, H | \theta_c)$
    $\Phi(\mathbf{x}^t) = (\gamma_1^t \Phi_1{}^T, \cdots, \gamma_c^t \Phi_c{}^T, \mathrm{KL}(\mathrm{Q}^t(\mathrm{y}) \| P(y)))^T$
end for
```

3.2 Computational Complexity Analysis

The computational complexity of the proposed discriminative extension is essentially similar to previous extensions [21]. In comparison with [21], the additional computational cost is (1) the estimation of α in the learning procedure (Algorithm 1) which takes only one step and is independent of the number of training samples N; and (2) the estimation of γ_c^t in the test procedure (Algorithm 2) which takes one step for each sample and is linearly dependent on the number of samples N_f .

4. Experimental Results

In this section, we apply our discriminative extension to three generative score space methods: FESS [8], PD [9] and SS [13], and refer to them as FESS-ours, PD-ours and SS-ours respectively. We cooperate with three generative models (GMMs [22], PSC [23] and LDA [7]) for image recognition. For each experiment, we run it on the randomly formed test and training sets for 10–20 rounds, and report the average accuracy.

4.1 Image recognition using Gaussian mixture model

In this experiment, we evaluate our method by cooperating with Gaussian mixture models (GMMs), which is a standard model for image representation, for image recognition. We choose three image datasets, Scene-15 dataset [22], OT dataset [24] and UIUC-sports dataset [25]. To further validate the effectiveness of our approach, we also choose a larger dataset, Caltech-101[26][27], for the image recognition task.

Deriving score functions from GMM. Here, we use Gaussian mixture models (GMMs) to model the distribution of image features for its effectiveness in image feature modeling and image recognition [28]. Suppose $\mathbf{x} \in \mathrm{R}^D$ be the observed variable (image feature), $\mathbf{z} = \{z_1, \cdots, z_K\}$ be a set of hidden variables (indicator) which follows Multinomial distribution over K possible events. The joint distribution of GMMs can be expressed as follows:

$$\begin{aligned} P(\mathbf{x},\mathbf{z};\theta) &= P(\mathbf{x}|\mathbf{z})P(\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x};\mathbf{u}_k,\Sigma_k)^{z_k} \prod_{k=1}^{K} \alpha_k^{z_k} \\ &= \prod_{k=1}^{K} \left[\frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}-\mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{u}_k) \right\} \right]^{z_k} \prod_{k=1}^{K} \alpha_k^{z_k} \end{aligned} \tag{23}$$

where $\mathbf{u}_k$ and Σ_k are the mean and covariance matrix of the k-th component; $\theta=\{\mathbf{u}_k,\Sigma_k\}_{k=1}^{K}$; $P(z_k)=\alpha_k$. The marginal distribution of GMMs is the integration of $P(\mathbf{x},\mathbf{z};\theta)$ over $\mathbf{z}$,

$$\begin{aligned} P(\mathbf{x};\theta) &= \sum_{z} P(\mathbf{x},\mathbf{z};\theta) \\ &= \prod_{k=1}^{K} \frac{\alpha_k}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}-\mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{u}_k) \right\} \end{aligned} \tag{24}$$

Let $Q^i(z) = \prod_{k=1}^{K} (g_k^i)^{z_k}$ be the posterior of hidden variables, conditioned on $\mathbf{x}^i$. As shown in Eq. (3), the free energy function for $\mathbf{x}^t$ can be written as

$$\begin{aligned} F(\theta) = -\sum_k g_k \log \frac{\alpha_k}{g_k} - \sum_k g_k (\log(2\pi)^{D/2} |\Sigma_k|^{1/2} \\ +\frac{1}{2} (\mathbf{x}-\mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{u}_k)) \end{aligned} \tag{25}$$

Having the lower bound of $\log P(\mathbf{x}^t|\theta)$, the elements of FESS score function are the summation terms of the lower bound and contain three groups,

$$\Phi(\mathbf{x}^i) = (\varphi_{fit}^{i,xT}, \varphi_{fit}^{i,zT}, \varphi_{ent}^{i,zT})^T \tag{26}$$

where

$$\begin{aligned} \varphi_{fit}^{i,x} &= (g_1^i \log N(\mathbf{x}^i;\mathbf{u}_1,\Sigma_1), \cdots, g_K^i \log N(\mathbf{x}^i;\mathbf{u}_K,\Sigma_K))^T \\ \varphi_{fit}^{i,z} &= (g_1^i \log a_1, \cdots, g_{1K}^i \log a_K)^T \\ \varphi_{ent} &= (g_1^i \log g_1^i, \cdots, g_K^i \log g_K^i)^T \end{aligned}$$

The score function for posterior divergence (PD) can be written as

$$\Phi^t = vec(\{\Phi_{x_d}^{pd}, \Phi_{x_d}^{fit}, \Phi_{z_k}^{pd}, \Phi_{z_k}^{fit}, \Phi_{z_k}^{ent}\}_{d,k}) \tag{27}$$

where

$$\begin{aligned} &\Phi_x^{pd} = \\ &\sum_{i\neq t}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} g_k^i \left(-\frac{(x_d^i - u_{d,+t})^2}{2\delta_{d,+t}^2} + \frac{(x_d^i - u_d)^2}{2\delta_d^2} - \log \frac{\delta_{d,+t}}{\delta_d} \right) \\ &\Phi_x^{fit} = \sum_{k,d=1}^{K,D} g_k^t \left(-\frac{(x_d^i - u_{d,+t})^2}{2\delta_{d,+t}^2} - \log \delta_{d,+t} (2\pi)^{\frac{D}{2}} \right) \\ &\Phi_z^{pd} = \sum_{i\neq t}^{N} \sum_{k=1}^{K} g_k^i \log \frac{\alpha_{k,+t}}{\alpha_k} \\ &\Phi_z^{fit} = \sum_{k=1}^{K} g_k^t \log \alpha_{k,+t} \end{aligned}$$

$$\Phi_z^{ent} = \sum_{k=1}^{K} g_k^t \log g_k^t$$

As shown in Section 3, we apply our proposed discriminative extension to FESS and PD. The resulting approaches are referred to as FESS-ours and PD-ours respectively.

Feature extraction. We use SIFT descriptor for image representation and dense sampling on a grid with the step size of 4 pixels. SIFT descriptors are extracted from three scales: 16×16，24×24 and 32×32. The k-means algorithm is used to form the codebook with 600 coding centers, and linear SVM is employed for classification. It is worth noting that the parameters except for the mixture centers K are learned from the joint model. We set $K = 60$ all throughout the experiments.

Experimental results on the UIUC-Sports dataset. The UIUC-Sports dataset [25] contains eight categories and 1792 images totally. The eight categories are *rock climbing, croquet, badminton, rowing, bocce, polo, sailing* and *snow boarding*. The number of images of each category ranges from 137 to 250. Some example images are shown in **Fig. 1**. As did in [29], we randomly select 70 images from each category for training and the rest for test. We repeat the experiment for 10 times and report the mean {and standard deviation}.

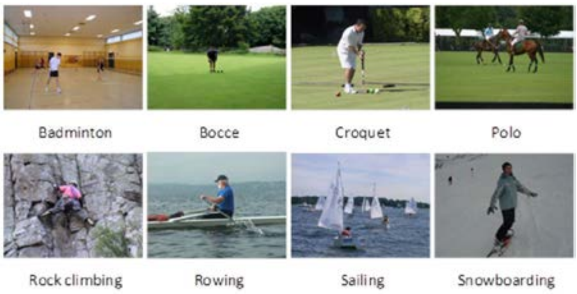


Fig. 1. Sample images from the UIUC dataset.

The proposed approaches, FESS-ours and PD-ours, will compare with several related methods (FESS, PD) and some state-of-the-art methods of image recognition task. ScSPM [30] uses sparse coding along with spatial pyramid matching. LScSPM [31] integrates the abilities of non-negative sparse coding, low-rank and sparse decomposition to form informative and robust representation. The adapted Gaussian mixture model (AGMM) [32] is a generic topic-independent Gaussian mixture model (known as the background GMM) is learned using all available training data and adapted to the individual topics. CSIFT locality-constrained linear coding (CLLC) [25] improves the performance of existing image classification algorithms by adding color information. The recognition accuracy of all methods on the UIUC-Sports dataset is shown in **Table 1**. As shown in **Table 1**, our proposed approach shows the best performance among all the compared approaches. Specifically, our discriminative extensions (FESS-ours, PD-ours) gain significant improvement over the existing generative score spaces (FESS, PD). The reason accounting for this superiority is that the proposed approach can exploit class label information, which encodes high-level information especially useful in image recognition. In fact, the proposed method benefits from the Bayes inference, where the posterior of class label is inferred and used as the components of score functions.

Table 1. The classification accuracy of our method by cooperating with GMM on the UIUC dataset.

Method	Accuracy (%)
ScSPM[30]	82.74±1.17
LScSPM[31]	85.31±0.51
AGMM[32]	82.50±1.89
CLLC[25]	82.98±1.23
FESS[8]	80.92±1.74
PD[9]	81.96±0.96
FESS-ours	85.86±1.57
PD-ours	**86.27±1.08**

Experimental Results on Scene-15 Dataset. The Scene-15 dataset [22] is composed of 15 scene categories. Each category comprises 200-400 images and there are totally 4485 medium size images. The scene categories vary from indoor scenes like kitchen and bedroom to outdoor scenes such as highway and street. Some example images are shown in **Fig. 2**. As were done in [22][33], we randomly choose 100 images from each category to form the training set and the rest serve as the test set. Like previous approaches, we report the mean {and standard deviation} after repeating the experiments 10 times.

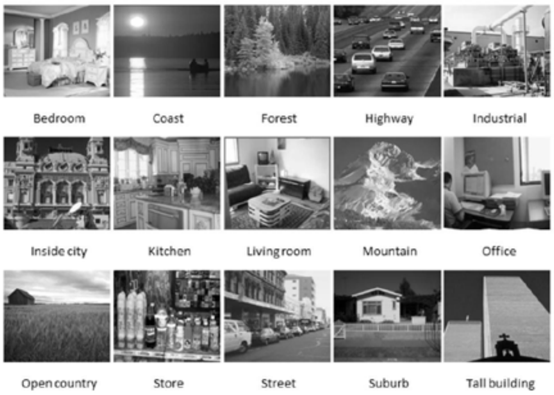


Fig. 2. Sample images from the Scene-15 dataset.

On Scene-15 dataset, our proposed discriminative extensions, denoted as FESS-ours and PD-ours, will compare with their initial versions (FESS, PD) closely related with our approaches and some state-of-the-art approaches, including bag-of-words (BoW) [34], AGMM [32], deep convolutional networks (DCN) [35], deep residual leaning (DRL) [36], and deep Fisher networks (DFN) [37]. BoW uses the histograms of visual words as the features of images, which is a baseline method without using feature learning. DCN [35] investigates the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. DRL [36] presents a residual learning framework to ease the training of networks that are substantially deeper than those used previously. DFN [37] proposes a version of the state-of-the-art Fisher vector image encoding that can be stacked in multiple layers.

Table 2. The classification accuracy of our method by cooperating with GMM on the Scene-15 dataset.

Method	Accuracy (%)
BoW[34]	79.30±1.64
AGMM[32]	83.20±0.96
DCN[35]	81.51±1.32
DRL[36]	80.22±1.21
FESS[8]	81.04±0.67
DFN[37]	81.48±1.55
PD[9]	82.87±1.21
FESS-ours	83.46±1.43
PD-ours	**83.92±1.18**

The recognition accuracy of our proposed method and other related methods on the Scene-15 dataset are shown in **Table 2**. As shown in **Table 2**, we find that, compared with the baseline approach BoW, AGMM, DCN, DRL obtain a significant improvement, and AGMM achieves the best performance among these three methods. Meanwhile, FESS, DFN and PD, which are three approaches most close to our proposed method, obtain competitive results due to the consideration of probabilistic modeling of data distribution. Our proposed approach, the discriminative extension for PD, achieves the best performance among the compared methods mentioned in **Table 2**, including FESS and PD. The reasons accounting for the improvement are: firstly, our method exploits class label which is very informative in image recognition; secondly, our method derives the feature mapping which encodes information over observed variables, hidden variables and model parameters.

Experimental Results on OT Dataset. The OT scene dataset [24] is composed of 4 natural scenes, *coast*, *forest*, *open country and mountain*, and 4 artificial scenes, *highway*, *inside city*, *street, and tall building*. There are 8 categories and 2688 images in total. The size of images is about 256×256 pixels. Sample images are presented in **Fig. 3**.

Fig. 3. Sample images from the OT dataset.

To further validate the effectiveness of the proposed approach, we then conduct an experiment on OT dataset. This dataset shares the same experimental setting as scene-15 dataset. We compared our proposed approach with FESS [8] and PD [9] closely related to our method and other state-of-the-art methods of this task, including BoW[34], DCN [35], DFN [37], CLLC [25], and AGMM[32]. Bag of words (BoW) [34] is a baseline method.

Table 3. The classification accuracy of our method by cooperating with GMM on the OT dataset.

Method	Accuracy (%)
BoW[34]	83.80±1.69
AGMM[32]	84.22±1.56
DCN [35]	83.98±1.58
CLLC[25]	84.65±1.09
FESS[8]	88.25±0.295
DFN[37]	85.05±1.92
PD[9]	87.98±1.15
FESS-ours	88.81±1.34
PD-ours	**90.01±0.76**

The recognition accuracy of all methods on the OT dataset are shown in **Table 3**. More specifically, we can see that AGMM, DCN and LCLA show competitive performance, both achieving a significant improvement over the baseline BoW. FESS, DFN and PD, which are closely related to our proposed approach, show competitive performance. This is because FESS, DFN and PD take probabilistic modeling of data distribution into account. Our proposed discriminative extension again outperforms other compared methods. The fact indicates that the proposed method exploits class label effectively, which encodes semantic information for image classification. Despite benefitting from generative information of image distribution when constructing feature mapping, the proposed approach also benefits from the Bayesian inference of class label. That is the reason why the approach is superior over other compared methods.

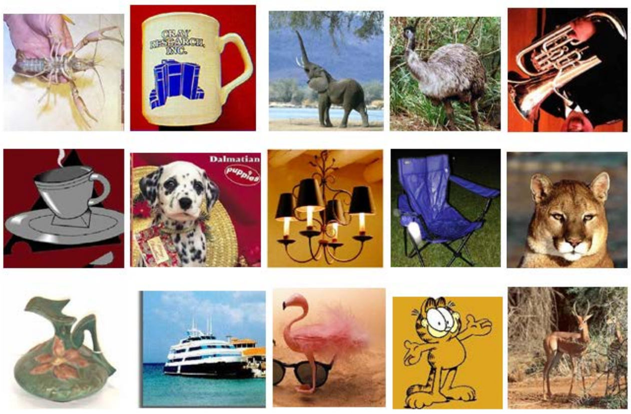


Fig. 4. Sample images from the Caltech-101 dataset.

Experimental Results on Caltech-101 Dataset. The Caltech-101 database is used for larger scale experiments, which contains 9,196 images. These images are classified into 101 categories, such as chair, barrel, anchor and dolphin, etc. The number of images varies along category in Catech-101 database. Sample images are presented in Fig.4. As were done in [38][39], we randomly choose 30 images from each category to form the training set and the rest serve as the test set. We repeat the experiment for 10 times and report the mean {and standard deviation}.

Our proposed approach will compare with several related approaches (FESS, PD) and some state-of-the-art methods of this task. Bag of features (BOF) [38] presents a method for recognizing scene categories based on approximate global geometric correspondence. Locally-constrained linear coding (LCLC) [39] proposes a fast approximated locality-constrained linear coding method by first performing a K-nearest-neighbor search and then solving a constrained least square fitting problem. ScSPM [30] uses sparse coding along with spatial pyramid matching. The results of BOF [38], LCLC [39] and ScSPM [30] on this dataset were previously reported in [26][40]. The experimental results are summarized in **Table 4**. Obviously, our approach (i.e. GMM+PD) again achieves the best performance among all the compared methods. The results validate the effectiveness of our approach on larger dataset. The reasons accounting for this are twofolds. First, the probabilistic generative modeling using GMM can exploit hidden information and is well adaptive to data distribution. Second, the performance of PD is proved that it can obtain recognition error rate as low as that of plug-in estimation. When we apply the extension to PD, the joint model can fully exploit class label information, which is very informative in classification.

Table 4. The classification accuracy of our method by cooperating with GMM on the Caltech-101 dataset.

Method	Accuracy (%)
BOF [38]	64.40±0.80
LCLC[39]	71.67±0.86
ScSPM[30]	72.20±1.30
FESS[8]	70.53±0.25
PD[9]	72.48±1.46
FESS-ours	72.91±1.09
PD-ours	**74.54±1.76**

4.2 Image recognition using probabilistic sparse coding

Probabilistic sparse coding (PSC) [23] assumes that samples are generated from the sparse and linear combination of overcomplete basis. For each sample, this approach aims to represent it using as fewer basis as possible. Suppose $\mathbf{x} \in R^D$ be observed variable; $\mathbf{z} \in R^K$ be hidden variable, i.e. the coefficients of linear combination, which follows Laplace distribution. We have

$$P(z_k) = \frac{1}{2b_k} \exp\left(-\frac{|z_k - u_k|}{b_k} \right) \tag{28}$$

where $u_k = 0, b_k = 1$. Given $\mathbf{z}$, the conditional distribution $P(\mathbf{x}|\mathbf{z})$ is Gaussian distribution. The joint probabilistic distribution of PSC can be written as:

$$P(\mathbf{x}, \mathbf{z}) = N(\mathbf{x}; \sum_k \mathbf{w}_k z_k, I) \prod_k \frac{1}{2} \exp(-|z_k|) \tag{29}$$

As were done in Section 2, we use the variational inference to derive score functions. The free energy function can be written as

$$F(Q,\theta) = \sum_{d,w} m(d,w) \sum_z Q(z|d,w) \log \frac{Q(z|d,w)}{P(d|z)P(w|z)P(z)} \tag{30}$$

The score function for posterior divergence (PD) can be written as

$$\Phi^t = vec(\{\Phi_{x_d}^{pd}, \Phi_{x_d}^{fit}, \Phi_{z_k}^{pd}, \Phi_{z_k}^{fit}, \Phi_{z_k}^{ent}\}_{d,k}) \tag{31}$$

where

$$\begin{aligned}\Phi_x^{pd} &= \sum_{i \neq t}^{N} E_{Q^i(\mathbf{z})} \log \frac{P(\mathbf{x}^i|\mathbf{z}, \theta_{+t})}{P(\mathbf{x}^i|\mathbf{z}, \theta_t)} \\ &= \sum_{i \neq t}^{N} \sum_{d=1}^{D} E_{Q^i(\mathbf{z})} \left[-\frac{(x_d^i - \sum_k z_k w_{d,k,+t})^2}{2\delta_{d,+t}^2} + \frac{(x_d^i - \sum_k z_k w_{d,k})^2}{2\delta_d^2} - \log \frac{\delta_{d,+t}}{\delta_d} \right], \\ &= \sum_{d=1}^{D} \Phi_{x_d}^{pd}\end{aligned}$$

$$\begin{aligned}\Phi_x^{fit} &= E_{Q^i(\mathbf{z})} \log P(\mathbf{x}^i|\mathbf{z}, \theta_{+t}) \\ &= \sum_{d=1}^{D} -\frac{(x_d^i - \sum_k z_k w_{d,k,+t})^2}{2\delta_{d,+t}^2} - \log \delta_{d,+t} (2\pi)^{\frac{\pi}{2}} , \\ &= \sum_{d=1}^{D} \Phi_{x_d}^{fit}\end{aligned}$$

$$\begin{aligned}\Phi_z^{pd} &= \sum_{i \neq t}^{N} E_{Q^i(\mathbf{z})} \log \frac{P(\mathbf{z}|\theta_{+t})}{P(\mathbf{z}|\theta)} , \\ &= \sum_{i \neq t}^{N} \sum_{k=1}^{K} 0 = \sum_{k=1}^{K} \Phi_{z_k}^{pd}\end{aligned}$$

$$\begin{aligned}\Phi_z^{fit} &= E_{Q^i(\mathbf{z})} \log P(\mathbf{z}|\theta_{+t}) \\ &= \sum_{k=1}^{K} E_{Q^i(\mathbf{z})} [-\log 2 - |z_k|] , \\ &= \sum_{k=1}^{K} \Phi_{z_k}^{fit}\end{aligned} \qquad \begin{aligned}\Phi_z^{ent} &= E_{Q^i(\mathbf{z})} \log Q^t(\mathbf{z}) \\ &= \sum_{k=1}^{K} E_{Q^i(\mathbf{z})} \left[-\log 2b_k^t - \frac{|z_k - u_k^t|}{b_k^t} \right]. \\ &= \sum_{k=1}^{K} \Phi_{z_k}^{ent}\end{aligned}$$

As shown in Section 3, we apply our proposed discriminative extension to PD. The resulting approaches are referred to as PD-ours.

Feature extraction. We use SIFT descriptor for image representation and dense sampling on a grid with the step size of 4 pixels. SIFT descriptors are extracted from three scales: 16×16，24×24 and 32×32. The k-means algorithm is used to form the codebook with 600 coding centers. Then each image is represented by a histogram of 600bins. Both support vector machine (SVM) [41] with linear kernel and localized multiple kernel learning (LMKL) [42] are chosen as classifiers.

Experimental results on the Scene-15 dataset. To validate the effectiveness of our proposed method PD-ours when cooperating with PSC, we compare them with related score space methods (FESS, PD, SS) and some state-of-the-art methods, including AGMM [32], DCN[35] and DRL[36]. We still use BoW as the baseline method. The recognition accuracy of all methods on the Scene-15 dataset are shown in **Table 5**. Our proposed discriminative extension (i.e. PSC+PD) with different classifiers all achieve excellent performance when compared with other score space methods and state-of-the-art methods. Especially, PD-ours +LMKL achieves the best performance. The reason accounting for this competitive performance is that the proposed method exploits class label information when deriving feature mapping. And the proposed method benefits from the Bayes inference when learning the feature mapping.

Table 5. The classification accuracy of our method by cooperating with PSC on the Scene-15 dataset.

Method	Classifier	Accuracy
	(%)	
BoW[31]	SVM	79.30±1.64
AGMM[29]	SVM	83.20±0.96
DCN [32]	SVM	81.51±1.32
DRL[33]	SVM	80.22±1.21
SS[13]	SVM	81.97±1.55
PD[9]	SVM	81.29±2.02
FESS[8]	SVM	80.76±1.41
PD-ours	SVM	82.87±1.43
SS[13]	LMKL	82.22±1.64
PD[9]	LMKL	81.95±1.87
FESS[8]	LMKL	82.28±2.14
PD-ours	LMKL	**83.17±1.23**

4.3 Image recognition using latent Dirichlet allocation

Latent Dirichlet allocation (LDA) [43] is proposed based on pLSA [7], which is a generative probabilistic model of a corpus. The difference between them is that LDA has hidden variables representing scene. The basic idea is that documents are represented as random mixtures over latent topics, where every topic is characterized by a distribution over words. Suppose a *word* be the basic unit of discrete data, indexed by $\{1,\cdots,V\}$; a *document* be a sequence of N words denoted by $\mathbf{w}=(w_1,w_2,\cdots,w_N)$, where w_n is the n-th word in the sequence; a *corpus* be a collection of M documents denoted by $D=\{\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_M\}$. LDA assumes the following generative process for each $\mathbf{w}$ in a corpus D:

(1) Choose N, which follows Poisson distribution parameterized by ξ;

(2) Choose θ, which follows Dirichlet distribution parameterized by α;

(3) For each of the N words w_n:

Choose a topic z_n, which follows Multinomial distribution parameterized by θ;

Choose a word w_n from $P(w_n | z_n, \beta)$, which is a multinomial probability conditioned on z_n.

Given the parameters α and β, The joint model of LDA over a topic mixture θ, a set of N topics $\mathbf{z}$ and a set of N words $\mathbf{w}$ can be written as:

$$P(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{k} (\alpha_i - 1) \prod_{n=1}^{N} P(z_n | \theta) P(w_n | z_n, \beta) \tag{32}$$

where $P(z_n | \theta)$ is simply θ_i for the unique i so that $z_n^i = 1$.

As were done in Section 2, we use the variational inference to derive score functions. The score function for SS can be characterized as follows.

$$\Phi(\mathbf{w}^d) = E_{Q(\mathbf{z}, \theta)}[\varphi(\mathbf{w}, \mathbf{z}, \theta)] \tag{33}$$

where $\varphi(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}) = vec\left(\{z_{dn}^k, w_{dn} z_{dn}^k, 1\}_{n,k}\right)$.

As shown in Section 3, we apply our proposed discriminative extension to SS. The resulting approaches are referred to as SS-ours. We conducted the experiment on OT dataset. The experiments are performed using two popular discriminative classifiers: support vector machine (SVM) [41] with linear kernel and localized multiple kernel learning (LMKL) [42]. We repeat each experiment for 20 rounds and report the average results, where in each round of experiment, we randomly select 30% samples for training and rest for test. The number of topic is determined by cross validation. We set the number of topics to be $K \in [40, 50]$. The classification results on the OT dataset are summarized in **Table 6**, where our discriminative extension outperform previous methods SS, FESS and PD over both SVM and LMKL.

Table 6. The classification accuracy of our method by cooperating with LDA on the OT dataset.

Method	#topic K	Classifier	Accuracy (%)
BOW[31]	---	SVM	83.80±1.69
AGMM[29]	---	SVM	84.22±1.56
DCN [32]	---	SVM	83.98±1.58
SS[13]	50	SVM	86.79±1.44
PD[9]	40	SVM	85.94±2.10
FESS[8]	40	SVM	87.65±1.45
SS-ours	50	SVM	89.98±1.65
SS[13]	50	LMKL	87.57±1.84
PD[9]	40	LMKL	86.43±3.02
FESS[8]	40	LMKL	87.92±1.99
SS-ours	50	LMKL	**91.35±1.23**

4.4 Discussion on the experimental results

We cooperate our discriminative extension with three generative models (GMM, PSC and LDA) for image recognition task. As shown in Table 1-Table 6, when comparing with closely related methods and state-of-the-art methods, our approach presents convincing results. The reasons accounting for this excellent performance can be summarized as follows. First, our method exploits class label which is very informative in image recognition; second, we model the joint distribution of the observed data and its class label, leading to a quite simple form, which can capture high-level relations within and across the class and visual modalities; third, the derived feature from generative model is essentially a function over hidden variable, model parameters, and observed data, which encodes high-level information important for image recognition. The computation cost of the proposed approach in real application includes three parts: (1) training the joint model; (2) applying the learned model for recognition; (3) the estimation of α in the learning procedure and γ_c^t in the test procedure. The first part is relatively time consuming because of the iteration of the EM algorithm. The second part is computationally effective since it only requires an E-step. The estimation in the third part only takes one step, and is independent of the number of training samples. Therefore, our discriminative extension can scale to large scale applications. Moreover, the performance of our approach can be potentially improved via exploiting spatial information. This idea will be left to future work.

5. Conclusion

In this paper, we propose a data-distribution-aware feature learning approach for content-based image recognition. The approach is based on exploiting class label from generative score space, in which image feature is derived from the probabilistic distribution of image datasets. The derived feature allows to fully exploit class label information and hidden information. Also, the derived score function is well adapt to data distribution. We cooperate our method with three generative models for image recognition task. The convincing experimental results demonstrate the effectiveness of our proposed extend generative feature learning approach. Our approach can be used under many practical conditions, such as image retrieval and sequence recognition. However, it still needs further evaluation to scale the method to other vision problems. Our method can further benefit from the mining of larger dataset.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.61503251), the Science and Technology Commission of Shanghai Municipality (No.16070502900), the Program of Shanghai Normal University (No. A-7001-15-001005), and the Natural Science Foundation of Shanghai (No. 15ZR1431500)

References

[1] Jiang Z, Zhang S, Zeng J. A, "hybrid generative/discriminative method for semi-supervised classification," *Knowledge-Based Systems*, vol. 37, no. 2, pp. 137-145, 2013. Article (CrossRef Link).

[2] B. Wang and Y. Liu, “Collaborative similarity metric learning for semantic image annotation and retrieval,” *Ksii Transactions on Internet and Information Systems*, vol. 7, no. 5, pp. 1252–1271, 2013. Article (CrossRef Link)

[3] C.Wang, B.Wang, and L. Zheng, “Learning free energy kernel for image retrieval,” *Ksii Transactions on Internet and Information Systems*, vol. 8, no. 8, pp. 2895–2912, 2014. Article (CrossRef Link)

[4] Zhou X, Jiang P, Wang X, “Recognition of control chart patterns using fuzzy SVM with a hybrid kernel function,” *Journal of Intelligent Manufacturing*, pp. 1-17, 2015. Article (CrossRef Link).

[5] Moran S, Lavrenko V, “A sparse kernel relevance model for automatic image annotation,” *International Journal of Multimedia Information Retrieval*, vol. 3, no. 4, pp. 209-229, 2014. Article (CrossRef Link).

[6] M. A. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1745 – 1752, 2011. Article (CrossRef Link).

[7] Bouguila N, “Hybrid Generative/Discriminative Approaches for Proportional Data Modeling and Classification,” *IEEE Transactions on Knowledge & Data Engineering*, vol. 24, no. 12, pp. 2184-2202, 2012. Article (CrossRef Link).

[8] A Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, “Free energy score spaces: using generative information in discriminative classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. Article (CrossRef Link)

[9] X. Li, S. L. Tai and Y. Liu, “Hybrid generative-discriminative classification using posterior divergence,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2713–2720, 2011. Article (CrossRef Link).

[10] T. Jebara, R. Kondor, A. Howard, “Probability product kernels,” *Journal of Machine Learning Research*, vol. 5, pp. 819-844, 2004. Article (CrossRef Link)

[11] N. Vasconcelos, “On the efficient evaluation of probabilistic similarity functions for image retrieval,” *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1482-1496, 2004. Article (CrossRef Link).

[12] Wang B, Wang C, Liu Y, “Exploiting class label in generative score spaces,” *Neurocomputing*, vol. 145, no. 18, pp. 495-504, 2014. Article (CrossRef Link).

[13] X. Li, B.Wang, Y. Liu, and S. L. Tai, “Stochastic feature mapping for pac-bayes classification,” *Machine Learning*, vol. 101, pp. 5–33, 2015. Article (CrossRef Link).

[14] B. Wang, X. Li and Y. Liu, “Learning discriminative fisher kernel for image retrieval,” *Ksii Transactions on Internet and Information Systems*, vol. 7, no. 3, pp. 532–548, 2013. Article (CrossRef Link)

[15] T. S. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” *Advances in Neural Information Processing Systems*, vol. 11, no. 11, pp. 487–493, 1998. Article (CrossRef Link)

[16] Amer M R, Siddiquie B, Tamrakar A, et al., “Human Social Interaction Modeling Using Temporal Deep Networks,” *Computer Science*, vol. 351, no. 2, pp. 193-197, 2015. Article (CrossRef Link)

[17] Wang B, Wang C, Huang J, “Multiple Clusters Parts-based Sparse Representation for Single Example Face Identification,” *Journal of Visual Communication & Image Representatio*n, vol. 40, pp. 237-250, 2016. Article (CrossRef Link).

[18] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods,” in *Proc. of British Machine Vision Conference*, pp. 76.1–76.12, 2011. Article (CrossRef Link).

[19] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, “Free-form region description with second-order pooling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1177–1189, 2015. Article (CrossRef Link).

[20] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Machine Learning*, vol. 37, no. 2, pp. 183–233, 2012. Article (CrossRef Link).

[21] K.Tsuda, M.Kawanabe, G.Ratsch, S.Sonnenburg, K.Muller, "A new discriminative kernel from probabilistic models," *Neural Computing*, vol. 14, no. 10, pp. 2397–2414, 2002. Article (CrossRef Link).

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Computer Science*, pp. 580–587, 2014. Article (CrossRef Link).

[23] Jin W, Ping L, Mary F H S, et al., "Biomedical time series clustering based on non-negative sparse coding and probabilistic topic model," *Computer Methods & Programs in Biomedicine*, vol. 111, no. 3, pp. 629-41, 2013. Article (CrossRef Link).

[24] Mennesson J, Saint-Jean C, Mascarilla L, "Color Fourier-Mellin Descriptors for Image Recognition," *Pattern Recognition Letters*, vol. 40, no. 1, pp. 27–35, 2014. Article (CrossRef Link).

[25] J. Chen, Q. Li, Q. Peng, and K. H.Wong, "Csift based locality-constrained linear coding for image classification," *Formal Pattern Analysis and Applications*, vol. 18, no. 2, pp. 441–450, 2015. Article (CrossRef Link).

[26] Zhang T, Ghanem B, Liu S, et al., "Low-Rank Sparse Coding for Image Classification," in *Proc. of IEEE International Conference on Computer Vision*, pp. 281-288, 2013. Article (CrossRef Link).

[27] Yang X, Zhang T, Xu C, "Locality discriminative coding for image classification," in *Proc. of International Conference on Internet Multimedia Computing and Service*, pp. 52-55, 2013. Article (CrossRef Link).

[28] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. of British Machine Vision Conference*, 2011. Article (CrossRef Link).

[29] B. Poczos, L. Xiong, D. J. Sutherland, and J. Schneider, "Nonparametric kernel estimators for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2989–2996, 2012. Article (CrossRef Link).

[30] J.Yang, K.Yu, Y.Gong, T.Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. Article (CrossRef Link)

[31] C.Zhang, J.Liu, Q.Tian, C.Xu, H.Lu, S.Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1673–1680, 2011. Article (CrossRef Link).

[32] M.Dixit, N.Rasiwasia, N.Vasconcelos, "Adapted Gaussian models for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.937–943, 2011. Article (CrossRef Link).

[33] Karpathy A, Fei-Fei L, "Deep visual-semantic alignments for generating image descriptions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 3128-3137, 2015. Article (CrossRef Link).

[34] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, vol. 11, no. 11, pp. 487–493, 2014. Article (CrossRef Link)

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Science*, 2015. Article (CrossRef Link)

[36] Simonyan K, Vedaldi A, Zisserman A, "Deep Fisher Networks for Large-Scale Image Classification," *Advances in Neural Information Processing Systems*, pp. 163-171, 2013. Article (CrossRef Link)

[37] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1673–1680, 2011. Article (CrossRef Link).

[38] Lazebnik S, Schmid C, Ponce J, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006. Article (CrossRef Link).

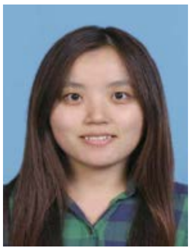
[39] Wang J, Yang J, Yu K, et al. "Locality-constrained Linear Coding for image classification," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010. Article (CrossRef Link).

[40] Boureau Y, Roux N L, Bach F, et al., "Ask the locals: Multi-way local pooling for image recognition," in *Proc. of IEEE International Conference on Computer Vision*, pp. 2651-2658, 2011. Article (CrossRef Link).

[41] V.Vapnik, "The Nature of Statistical Learning Theory," Springer, Verlag, New York, 2000. Article (CrossRef Link).

[42] M.Gönen, E.Alpaydin, "Localized multiple kernel learning," in *Proc. of International Conference on Machine Learning*, pp. 352–359, 2008. Article (CrossRef Link)

[43] Blei D M, Ng A Y, Jordan M I, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003. Article (CrossRef Link)



Bin Wang received her PhD degree from Department of Automation, Shanghai Jiao Tong University, Shanghai, China. She is now the lecturer of Shanghai normal University, China. Her research interests include computer vision, machine learning, image processing, multimedia analysis.

Chuanjiang Li received his Ph. D. degree from Shanghai University in 2014. He is currently an associate professor in Shanghai Normal University. His current research interests include intelligent mobile robots, human robot interaction.

Qian Zhang is now the lecturer of Shanghai normal University, China. She received her P.HD. from Shanghai University in China. Her research interest fields include video processing.

Jifeng Huang received the Ph.D. degree from East China University of Science and Technology, Shanghai, China, in 2005. He has been a professor in Shanghai Normal University from 2007. His research interests are in image processing and computer vision, especially in image processing.