Generative and Discriminative Sparse Coding for Image Classification Applications

Ajit Puthenputhussery Qingfeng Liu Hao Liu Chengjun Liu
New Jersey Institute of Technology
avp38, ql69, hl422, cliu@njit.edu

Abstract

This paper presents an enhanced sparse coding method by exploiting both the generative and discriminative information in sparse representation model. Specifically, the proposed generative and discriminative sparse representation (GDSR) method integrates two new criteria, namely a discriminative criterion and a generative criterion, into the conventional sparse representation criterion. The generative criterion reveals the class conditional probability of each dictionary item by using the dictionary distribution coefficients which are derived by representing each dictionary item as a linear combination of the training samples. To further enhance the discriminative ability of the proposed method, a discriminative criterion is also applied using new localized within-class and between-class scatter matrices. Moreover, a novel GDSR based classification (GDSRc) method is proposed by utilizing both the derived sparse representation and the dictionary distribution coefficients. This hybrid method provides new insights, and leads to an effective representation and classification schema for improving the classification performance. The largest step size for learning the sparse representation is theoretically derived to address the convergence issues in the optimization procedure of the GDSR method. Extensive experimental results and analysis on several public classification datasets show the feasibility and effectiveness of the proposed method.

1. Introduction

In recent years, machine learning and computer vision techniques have been broadly applied for several classification tasks such as object classification [27, 33, 3, 31, 29], scene classification [20, 21, 36, 4, 5], face recognition [33, 41, 9, 39, 34, 22], and fine grained classification [17, 18, 7]. However, in order to accurately classify images, a discriminative and robust representation is needed to capture the important aspects of the image. A major issue in computer vision applications is the high dimensionality of the image feature vector which can make the learning tasks more difficult and can have a dramatic impact on the performance. To solve this issue, sparse coding algorithms [19, 35, 34, 28] have been widely used for data modeling by learning a dictionary that is adapted to the data to improve the feature representation. Sparse coding allows efficient retrieval of data as it generates sparse representations such that every data point can be represented as a linear combination of a small set of basis vectors. Another advantage is that the sparse representation can be overcomplete, allowing more flexibility in matching data and yielding a better approximation of the statistical distribution of the data.

Although the sparse representation method achieves impressive results in various challenging tasks, a potential limitation is the lack of generative information since the dictionary is only derived from the representation criterion. The generative perspective remains ignored due to the intrinsic difficulty of estimating the class conditional probability accurately. The generative criterion models the data distribution and infers joint representations which may significantly affect the performance of the learning system. Another limitation in the conventional sparse representation criterion is the lack of discriminative criterion which helps to enhance the discrimination among data samples of different categories. Previous works of research by [25, 8] show the complementary nature of discriminative and generative approaches and demonstrate the effectiveness of combining both the approaches.

To address these limitations, we present a novel generative and discriminative sparse representation (GDSR) method by integrating the conventional sparse representation, a generative criterion and a discriminative criterion. The proposed GDSR method intrinsically models a hybrid paradigm of both the generative information and the discriminative information. It also helps to avoid over-fitting as the generative model acts as a regularizer for the discriminative model from the regularization point of view. Figure 1 illustrates the framework of the proposed GDSR method. Specifically, the generative criterion plays the role of generative modeling by representing each dictionary item as

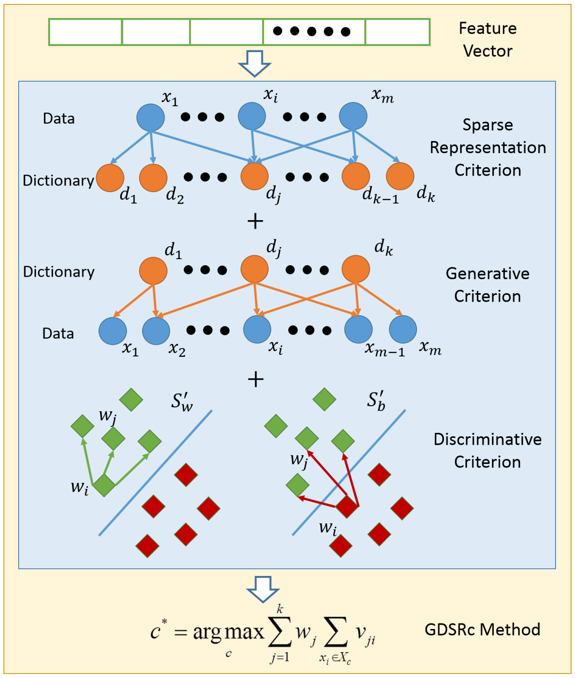


Figure 1. The framework of the proposed GDSR method.

a linear combination of the training samples and also emphasizes the coefficients of the nearest training samples. Theoretical analysis shows that these coefficients known as the "dictionary distribution coefficients", are capable of approximately modeling the class conditional probability of each dictionary item. To further improve the classification capability and utilize the marginal information, a discriminative criterion is integrated that applies newly defined within-class and between-class scatter matrices by considering only the k nearest neighbors. We then propose a new classification method, namely the generative and discriminative sparse representation based classification (GDSRc) method, that exploits both the new sparse representation and the dictionary distribution coefficients. Finally, we theoretically derive the largest step size for learning the sparse representation to address the convergence issues of our proposed optimization procedure. Our proposed GDSR method is evaluated on several publicly available classification datasets and the experimental results shows that GDSR method achieves better results compared to other sparse coding and learning methods.

2. Related Work

In image classification applications, several machine learning methods such as sparse coding, deep learning and manifold learning have been widely used to develop an effective and robust representation schema to improve the classification performance. A locally linear KNN (LLKNN) model was developed by Liu et al. [21] that derives a new representation by using the criteria of locality and sparsity. Jindal et al. [10] proposed a method to train a deep network from training data having noisy labels by augmenting the deep network with a softmax layer that models the label noise.

Recently, three categories of dictionary learning methods have been proposed for sparse representation. The first category co-trains the discriminative dictionary, sparse representation and the linear classifier together. Mairal et al. [24] proposed to co-train the discriminative dictionary, sparse representation as well as the linear classifier using a combined objective function. The D-KSVD method was developed by Zhang et al. [41] to learn the discriminative dictionary and the classifier simultaneously. Jiang et al. [9] improved upon the method introduced in [41] by introducing a label consistent regularization term.

The second category combines the sub-dictionaries to utilize their discriminative power. Zhou et al. [44] presented a Joint Dictionary Learning (JDL) method that jointly learns both the commonly shared dictionary and the class-specific sub-dictionaries to exploit the correlation between similar data samples. A dictionary learning approach was developed by Yang et al. [39] that learns a structured dictionary containing a set of class-specific sub-dictionaries.

The third category learns the dictionary by modeling the relation between the dictionary and each class label. Yang et al. [37] proposed a latent dictionary learning (LDL) method by jointly learning a latent vector which indicates the relation between the dictionary and the labels. A discriminative Bayesian dictionary learning (DBDL) method was developed by Naveed et al. [1] that infers the distribution of the dictionary using an approximation of the Beta process.

Our method differs from the above methods in the following aspects. First, our method exploits both the generative information and the discriminative information in the sparse representation model in comparison with other methods. Second, our method does not depend on any assumption about the probability distribution, such as Bernoulli distributions in DBDL [1]. And finally, our method does not depend on the sub-dictionary, which might lead to overfitting and deteriorate the performance when the training data of each class is not sufficient.

3. Generative and Discriminative sparse representation (GDSR)

In this section we derive a novel sparse representation model by exploiting both the generative and the discriminative information to improve the classification performance. Dictionary learning plays a crucial role in the conventional sparse representation method. An important question that arises during the dictionary learning process, which receives much less explicit attention, is how a dictionary item is generated given a specific category, namely the generative information of the dictionary. One naive answer is to construct a dictionary that consists of carefully selected training samples [34], [21]. In this scenario, each dictionary item corresponds to a training sample from a specific category. Such a dictionary might achieve good results, however, the performance of this method relies heavily on the selection of the training samples and the size of the selected training samples.

Our solution to the above question is the proposed GDSR method, which explicitly models the class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$, where $\mathbf{d}_j$ is j-th the dictionary item and c is the class label, and introduces a new discriminative criterion for enhancing the discriminative power of the dictionary. Given the training sample data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ that contains m samples $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$, and each sample resides in the n dimensional space. The dictionary $\mathbf{D} \in \mathbb{R}^{n\times k}$ can be represented as $[\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k]$, where each dictionary item $\mathbf{d}_j (j = 1, 2, ..., k)$ also resides in the n dimensional space. Then our GDSR method derives the sparse representation $\mathbf{w}_i \in \mathbb{R}^{k\times 1} (i = 1, 2, ..., m)$ for each training sample $\mathbf{x}_i$, and the dictionary distribution coefficients $\mathbf{v}_j \in \mathbb{R}^{m\times 1} (j = 1, 2, ..., k)$ for each dictionary item $\mathbf{d}_j$.

Specifically, the GDSR method is defined as follows:

$$\min_{\mathbf{D},\mathbf{W},\mathbf{V}} \{\sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \lambda||\mathbf{w}_i||_1\} + \gamma L(\mathbf{V}, \mathbf{D}) + \alpha H(\mathbf{W})$$
$$s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \tag{1}$$

The first term in equation 1 is the conventional sparse representation criterion, where the parameter λ controls the L_1 normalization.

The second term $L(\mathbf{V}, \mathbf{D})$ is the generative criterion, which is defined as follows:

$$L(\mathbf{V}, \mathbf{D}) = \sum_{j=1}^{k} ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j||^2 + \sigma||\mathbf{v}_j - \eta\mathbf{p}_j||^2 \tag{2}$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k]$ is the matrix that consists of the dictionary distribution coefficients vector $\mathbf{v}_j = [v_{j1}, v_{j2}, ..., v_{jm}]^t$. The vector $\mathbf{p}_j = [p_{j1}, p_{j2}, ..., p_{jm}]^t \in \mathbb{R}^m$ represents the distance measure between the dictionary item $\mathbf{d}_j$ and the training sample $\mathbf{x}_i$ as follows:

$$p_{ji} = exp\{-\frac{1}{2h^2}||\mathbf{d}_j - \mathbf{x}_i||^2\} \tag{3}$$

where the parameter h controls the decay speed. Note that $p_{ji} \leq 1$ and $||\mathbf{p}_j||^2$ can be normalized.

The traditional view of the dictionary learning is to represent the training sample as a linear combination of the dictionary items. In comparison as shown in figure 1, our generative criterion demonstrates a reciprocal viewpoint as well, which represents each dictionary item as a linear combination of the training samples. The dictionary items and the training samples consist of a bipartite graph and they influence each other mutually. In addition, the generative criterion also adds a constraint on the dictionary distribution coefficients vector $\mathbf{v}_j$ such that the coefficients are proportional to the distance between the dictionary item and the training sample, in order to estimate the class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$ by using $\mathbf{v}_j$ (Proposition 3.1).

The third term is the discriminative criterion, which is defined as follows:

$$H(\mathbf{W}) = \mathbf{tr}(\beta\mathbf{S}_w^{'} - (1 - \beta)\mathbf{S}_b^{'}) \tag{4}$$

where the new within-class scatter matrix is defined as $\mathbf{S}_w^{'} = \sum_{i=1}^{m} \sum_{(\mathbf{w}_i,\mathbf{w}_j)\in T_k^w} (\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^t$, and T_k^w represents the set of $(\mathbf{w}_i, \mathbf{w}_j)$ pairs where the sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$ are among their k nearest neighbors respectively in the same class. The new between-class scatter matrix is defined as $\mathbf{S}_b^{'} = \sum_{i=1}^{m} \sum_{(\mathbf{w}_i,\mathbf{w}_j)\in T_k^b} (\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^t$, where T_k^b represents the set of the k nearest $(\mathbf{w}_i, \mathbf{w}_j)$ pairs among all the $(\mathbf{w}_i, \mathbf{w}_j)$ pairs between sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$ from different classes.

This discriminative criterion utilizes the underlying topology of the sparse representation of training samples for defining new within-class and between-class scatter matrices by considering only the k nearest neighbors. The new discriminative criterion can be further transformed to $H(\mathbf{W}) = \mathbf{tr}(\mathbf{W}\mathbf{L}\mathbf{W}^t)$, where $\mathbf{L} = 2\beta(\mathbf{D}_w - \mathbf{W}_w) - 2(1 - \beta)(\mathbf{D}_b - \mathbf{W}_b)$. In particular, let $\mathbf{W}_w$ be a matrix, whose elements $W_w(i, j) = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are among the k nearest neighbors of each other in the same class, and $W_w(i, j) = 0$ otherwise. Let $\mathbf{W}_b$ be a matrix, whose elements $W_b(i, j) = 1$ if the pair $(\mathbf{w}_i, \mathbf{w}_j)$ is among the k nearest pairs from all the pairs among the samples of different classes, and $W_b(i, j) = 0$ otherwise. And, let $\mathbf{D}_w$ and $\mathbf{D}_b$ be diagonal matrices, whose main diagonal elements are $D_w(i, i) = \sum_{j\neq i} W_w(i, j)$, and $D_b(i, i) = \sum_{j\neq i} W_b(i, j)$, respectively. An important property of the proposed GDSR method is the modeling of class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$ stated as the following generative property 3.1.

Proposition 3.1. ***Generative Property*** *Given that* $\mathbf{V}$ *is the derived dictionary distribution coefficients by the proposed GDSR method, the class conditional probability of each dictionary item* $p(\boldsymbol{d}_j|c)$ *is modeled as follows.*

$$p(\boldsymbol{d}_j|c) \propto \sum_{\boldsymbol{x}_i\in\boldsymbol{X}_c} v_{ji} \tag{5}$$

where $\mathbf{X}_c$ *is the set of training samples in the c-th class.*

The conventional way to estimate $p(\mathbf{d}_j|c)$ assumes some parametric distribution first, such as the Bernoulli distribution. But in comparison, the generative property of our proposed method shows that $p(\mathbf{d}_j|c)$ is estimated from the kernel density estimation point of view. Our GDSR method provides a coarse estimation of the class conditional probability of each dictionary item instead of an accurate estimation, since our goal is to correctly classify the data instead of accurately estimating the probability. Such a coarse modeling carries sufficient information for improving the classification performance as shown in the Experiments section.

4. Optimization Procedure

In this section, we discuss the optimization procedure of the proposed GDSR method. The objective function in equation 1 is optimized using a coordinate descent method, which alternatively updates the sparse representation, the dictionary distribution coefficients, as well as the discriminative dictionary. In order to obtain a better convergence rate, the sparse representation and the dictionary are initialized using the conventional sparse representation method [13], while the dictionary distribution coefficients $\mathbf{v}_j$ are initialized using the value of $\eta\mathbf{p}_j$.

4.1. Updating the Sparse Representation

First, given the dictionary $\mathbf{D}$ and the dictionary distribution coefficients $\mathbf{V}$, the sparse representation $\mathbf{W}$ for each training sample $\mathbf{x}_i$ can be obtained by rewriting the objective function defined in equation 1 as follows.

$$\min_{\mathbf{w}_i} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \alpha L_{ii}\mathbf{w}_i^t\mathbf{w}_i + \alpha\mathbf{w}_i^t\mathbf{h}_i + \lambda||\mathbf{w}_i||_1; \quad (6)$$

where $\mathbf{h}_i = \sum_{j\neq i} L_{ij}\mathbf{w}_j = [h_{i1}, h_{i2}, ..., h_{ik}]^t$ and $L_{ij}(i,j = 1, 2, ..., m)$ is the value in the i-th row, j-th column of the matrix $\mathbf{L}$. We then apply the FISTA algorithm [2] to learn the sparse representation $\mathbf{w}_i$ for each training sample $\mathbf{x}_i$.

To guarantee the convergence of the FISTA algorithm, an important quantity to be determined is the step size. Given the objective function $F(x) = f(x) + g(x)$, where $f(x)$ is a smooth convex function and $g(x)$ is a non-smooth convex function, the theoretical analysis [2] shows that

$$F(x_k) - F(x^*) \leq \frac{2||x_0 - x^*||^2}{s * (k+1)^2} \quad (7)$$

where x_k is the solution generated by the FISTA algorithm at the k-th iteration, x^* is the optimal solution, and s is the largest step size for convergence. This theoretical result means that the number of iterations of the FISTA algorithm required to obtain an ϵ-optimal solution (x_t), such that $F(x_t) - F(x^*) \leq \epsilon$, is at most $\lceil C/\sqrt{\epsilon} - 1\rceil$, where $C = \sqrt{2||x_0 - x^*||^2/s}$ Therefore, the step size plays an important role for the convergence of the algorithm and the largest step size can lead to less required iterations for the convergence of the FISTA algorithm. The largest step size required for learning the sparse representation for each training sample is stated in Proposition 4.1.

Proposition 4.1. *The largest step size that guarantees convergence of the FISTA algorithm is* $\frac{1}{Lip(f)}$, *where* $Lip(f)$ *is the smallest Lipschitz constant of the gradient* ∇f *and* $Lip(f) = 2E_{\max}(\boldsymbol{D}^t\boldsymbol{D} + \alpha L_{ii}\boldsymbol{I})$ *which is twice the largest eigenvalue of the matrix* $(\boldsymbol{D}^t\boldsymbol{D} + \alpha L_{ii}\boldsymbol{I})$.

4.2. Updating the Dictionary Distribution Coefficients

Second, when the dictionary $\mathbf{D}$ and the sparse representation $\mathbf{W}$ are given, the dictionary distribution coefficients $\mathbf{V}$ can be derived using the following analytical solution.

$$\mathbf{v}_j = (\mathbf{X}^t\mathbf{X} + \sigma\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{d}_j + \sigma\eta\mathbf{p}_j) \quad (8)$$

where $\mathbf{X}^t\mathbf{d}_j$ is the sample correlation between the dictionary item $\mathbf{d}_j$ and all the training samples, and $\mathbf{p}_j$ is the reciprocal of the exponential form of Euclidean distance between $\mathbf{d}_j$ and all the training samples. Therefore, the dictionary distribution coefficient $\mathbf{v}_j$ represents a measurement between the dictionary item and the training samples using a combination of both the correlation information and the distance information. From another perspective, $\mathbf{v}_j$ is a similarity measure using both the angular distance (correlation information) and the Euclidean distance (reciprocal of the exponential form of Euclidean distance). This important property of $\mathbf{v}_j$ significantly helps to derive the dictionary as shown in the following sub-section.

4.3. Updating the Dictionary

Third, after learning the sparse representation $\mathbf{W}$ and the dictionary distribution coefficients $\mathbf{V}$, the dictionary $\mathbf{D}$ can be derived by optimizing the following objective function.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{D}\mathbf{W}||^2 + \gamma(||\mathbf{D} - \mathbf{X}\mathbf{V}||^2 + \sigma||\mathbf{V} - \eta\mathbf{P}||^2) \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \quad (9)$$

where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k]$. The optimization of equation 9 is not a trivial problem due to the exponential form of the vector $\mathbf{p}_j$ with respect to $\mathbf{d}_j$. We seek a more efficient approximation to derive the dictionary instead of using some generic solvers. It is based on the observation from equation 8 that the coefficients of the nearest neighbors of the dictionary items are sufficient for an efficient approximation since the dictionary distribution coefficient vector $\mathbf{v}_j$ represents a similarity measure between the training samples and the dictionary items. Specifically, the approximation method consists of the following steps. (i) The influence of distant training samples are diminished by setting

Dataset	Task	# Classes	# Images
15 Scenes [12]	scene classification	15	4485
MIT-67 [30]	scene classification	67	15620
Caltech 256 [6]	object classification	256	30607
Extended Yale B [14]	face recognition	38	2414

Table 1. Description of the different data sets used for evaluation of the proposed GDSR method.

the elements whose absolute value is less than a threshold in $\mathbf{v}_j$ to zero. The resulting new vector is denoted as $\bar{\mathbf{v}}_j$. (ii) The dictionary is then derived by solving the following new optimization problem.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X}-\mathbf{D}\mathbf{W}||^2 + \gamma||\mathbf{D}-\mathbf{X}\bar{\mathbf{V}}||^2 \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{10}$$

where $\bar{\mathbf{V}}$ is a matrix containing $\bar{\mathbf{v}}_j$. This problem is a constrained optimization problem with inequality constraints, which is solved using the Lagrange optimization and the Karush-Kuhn-Tucker condition [13].

4.4. Generative and Discriminative Sparse Representation based Classification (GDSRc)

After the dictionary $\mathbf{D}$ and the dictionary distribution coefficients $\mathbf{V}$ are derived, we present a new generative and discriminative sparse representation based classification (GDSRc) method. In particular, for the test data $\mathbf{y}$, we derive sparse representation by optimizing the following criterion:

$$\min_{\mathbf{w}} \left\{ ||\mathbf{y}-\mathbf{D}\mathbf{w}||^2 + \lambda||\mathbf{w}||_1 \right\} \tag{11}$$

where the representation $\mathbf{w} = [w_1, w_2, ..., w_k]^t$ contains both the generative and the discriminative information, as the dictionary $\mathbf{D}$ is learned during the training optimization process.

The GDSRc method is then applied based on the derived generative and discriminative sparse representation $\mathbf{w}$ and the dictionary distribution coefficients $\mathbf{v}$. Specifically, the GDSRc method is defined as follows.

$$c^* = \arg\max_{c} \sum_{j=1}^{k} w_j \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} \tag{12}$$

Note that we only select the top T largest values of v_{ji} for the GDSRc method.

5. Experiments

We evaluate the effectiveness of the proposed GDSR method on several publicly available classification datasets listed in table 1. In order to have a fair comparison, we follow the same experimental protocol used by other methods for the respective datasets. Besides, we also preFacesent additional comprehensive analysis to understand the properties and the effect of the proposed GDSR method in this section. Note that the parameters for the dictionary distribution criterion are selected as $\gamma = 0.05$, $\sigma = 0.05$, and $\eta = 0.1$ for all the datasets. The parameters for the GDSR method are selected based on a grid search with cross validation approach. The metric used for performance evaluation is the classification accuracy.

Methods	Accuracy %
ROI + Gist [30]	26.1
DPM [26]	30.4
Object Bank [15]	37.6
miSVM [16]	46.4
D-Parts [32]	51.4
LLNMC [20]	59.12
DP + IFV [11]	60.8
Hybrid-CNN [43]	70.80
VGG16-Place365 CNN [42]	76.53
DAG-CNN [40]	77.50
GDSR	**82.97**

Table 2. Comparison with the other state-of-the-art methods on the MIT-67 indoor scenes dataset.

5.1. The MIT-67 Indoor Scenes Dataset

The MIT-67 indoor scenes dataset [30] is a challenging indoor scene classification dataset, which contains 67 indoor categories. We follow the experimental settings defined in [30] where 80 data samples per category are used for training and 20 data samples per category are used for testing. The initial input features used are extracted from a pretrained VGG16 CNN model [42] and the feature dimension is reduced from 4096 to 3500. For the GDSR method, the model parameters are selected as follows: $\lambda = 0.05$, $h = 0.01$, $\alpha = 0.1$, and $\beta = 0.5$. The dictionary size is set as 2048 and k is set as 75 for the GDSRc method.

We compare our proposed GDSR method with different sparse coding as well as deep learning methods. In particular, our proposed GDSR method helps to significantly improve the initial CNN input features (VGG16-Place365 CNN) by encouraging better separation between the samples of different class and assist in the formation of compact clusters for the samples of same class (see subsection 5.5). Experimental results in table 2 show that the proposed method is able to achieve significantly better results compared to popular sparse coding and deep learning methods since it uses both the generative and discriminative information.

Methods	Accuracy %
LLC [33]	89.20
D-KSVD [41]	89.10
LC-KSVD1 [9]	90.40
LC-KSVD2 [9]	92.90
LaplacianSC [4]	89.7
DHVFC [5]	86.4
Places-CNN [43]	90.19
Hybrid-CNN [43]	91.59
VGG16-Place365 CNN [42]	92.15
DAG-CNN [40]	92.90
GDSR	**98.90**

Table 3. Comparison with the other state-of-the-art methods on the 15 scenes dataset.

Methods	30	45	60
ScSPM [36]	34.02	37.46	40.14
IFK [27]	40.80	45.00	47.90
LLC [33]	41.19	45.31	47.68
M-HMP [3]	48.00	51.90	55.20
ZFNet CNN [31]	70.60	72.70	74.20
GDSR	**72.39**	**75.13**	**76.90**

Table 4. Comparison between the proposed method and other popular methods on the Caltech 256 dataset.

5.2. The 15 Scenes Dataset

The 15 scenes dataset [12] contains 4485 images from 15 scene categories, each with the number of images ranging from 200 to 400. The experimental protocol used for comparison is defined in [12] where 100 images per class are randomly selected for training and the remaining for testing for 10 iterations. The input features used are spatial pyramid features [9] obtained by using a four-level spatial pyramid and a codebook of size 200 resulting in a feature vector of dimension 3000 which is further reduced to 1000 using PCA dimensionality reduction method. We select the model parameters as follows: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, and $\beta = 0.5$. The size of the dictionary is set as 1024, and k=100 for the GDSRc method. The results shown in table 3 demonstrate that the proposed method is able to achieve better results compared to other learning methods.

5.3. Caltech 256 Data Set

The Caltech 256 dataset [6] is an extended version of the Caltech 101 dataset and a more challenging object classification dataset containing 30607 images from 256 categories. We follow the experimental protocol defined in [33] where the entire dataset is partitioned randomly into 15, 30, 45 and 60 training data samples per category and at the most 25 test data samples per category for 3 iterations. The initial input features used are extracted from a pre-trained ZFNet [31] resulting in feature vector with dimension 4096. We further reduce the dimension to 2000 using PCA. The metric used for performance evaluation is the average classification accuracy over all the categories. For the GDSR method, we set the dictionary size to 1024, and the parameters as λ = 0.05, $h = 0.1$, α = 0.1, β = 0.5, and k=60 for the GDSRc method. The experimental results in table 4 show that our proposed method is able to achieve better results compared to other methods.

Experimental setting 1	Accuracy %
D-KSVD [41]	75.30
SRC [34]	90.00
FDDL [39]	91.90
GDSR	**95.19**
Experimental setting 2	Accuracy %
LLC [33]	90.70
D-KSVD [41]	94.79 ± 0.49
LC-KSVD1 [9]	93.59 ± 0.54
LC-KSVD2 [9]	95.22 ± 0.61
FDDL [39]	96.07 ± 0.64
SRC [34]	96.32 ± 0.85
DBDL + SVM [1]	96.10 ± 0.25
GDSR	**97.45** ± 0.40

Table 5. Comparison with the state-of-the-art methods on the extended Yale face database B under two experimental settings.

5.4. Extended Yale face database B

The extended Yale face database B consists of 2414 frontal view face images from 38 individuals each with around 64 images taken under various lighting conditions. We use a cropped version of the database [14], where all the images are aligned and re-sized to 168 × 192. To show the robustness of our proposed method, we present results of our GDSR method under an extremely noisy condition using random faces [34] as the input features. Specifically, the random faces [34] consists of the row vectors of a randomly generated transformation matrix from a zero-mean normal distribution, which is applied to project the face pattern vector into a dimension of 504 representation vector.

We follow two common experimental settings to have a fair comparison with other methods. The first experimental setting is defined in [38] where 20 data samples per subject are randomly selected for training, and the remaining data samples are used for testing for 10 iterations. The model parameters for the GDSR method are selected as follows: $\lambda = 0.1$, $h = 0.1$, $\alpha = 0.5$, $\beta = 0.5$, $k = 20$ and the dictionary size is set to 512.

The second experimental setting is defined in [1, 9] where 32 data samples are randomly selected for training for each subject, and the remaining data samples are se-

Method	Accuracy (%)
GDSR with only discriminative criterion	77.24
GDSR with only generative criterion	78.51
Proposed GDSR (both criteria)	**80.67**

Table 6. Evaluation of the contribution of generative and discriminative criterion in GDSR method using the MIT-67 scenes dataset.

lected for testing for 10 iterations. The dictionary size is 512, and the model parameters are set as $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, $\beta = 0.5$, and $k = 32$ for GDSRc method. The results shown in table 5 demonstrate the effectiveness of the proposed method.

5.5. Evaluation of the Effect of the Proposed GDSR Method

To evaluate the contribution of the individual criterion to the overall classification accuracy, we conduct experiments on the MIT-67 dataset using the initial input features as described in the Experiments section 5.1. In order to have a fair comparison, we use the RBF-SVM classifier for classification instead of the GDSRc method since it depends on both the generative and discriminative criteria. It can be seen from table 6 that the GDSR method (both discriminative and generative criteria) achieves the best performance of 80.67% since it incorporates both the discriminative and the generative information.

We further discuss the effects of our proposed method on the initial features and how it encourages better clustering and discrimination among different classes of a dataset. To visualize the effect of our proposed method, we use the popular t-SNE visualization technique [23] that produces visualization of high dimensional data in scatter plots. Figure 2 shows the t-SNE visualizations of the initial features used as input and the features extracted after applying the GDSR method for different datasets. It can be seen from figure 2 that the proposed GDSR method helps to reduce the distance between images of the same class leading to formation of higher density clusters for images of the same class. Another advantage is that the GDSR method assists to increase the distance between clusters of different classes resulting in better discrimination among them. The GDSR method uses both the generative and discriminative information, therefore, encourages better separation between data samples of different classes.

5.6. Evaluation of the Feature Dimensionality and the Dictionary Size

This section presents an analysis of the performance under different sizes of dictionary and different values of feature dimensionality on the MIT-67 indoor scenes dataset [30]. Specifically, the dictionary sizes 128, 256, 512, 1024, and 2048 are evaluated for dimensionality 500, 1000, 1500, 2000, 2500, 3000, 3500, and 4000, respectively. From Figure 3, we can conclude that (i) larger dictionary sizes usually result in better performance, (ii) the low dimension feature space is more sensitive to over-completeness (dictionary size larger than the dimensionality) of the dictionary compared to the higher dimension feature space. For example, the dimensionality 500 requires at least a dictionary size of 512 to achieve good performance (above 80%), whereas a higher dimensionality feature of dimension 3500 only requires a dictionary size of 256.

Methods	Accuracy %
KNN	76.70
Linear-SVM	79.60
RBF-SVM	80.67
GDSRc	**82.97**

Table 7. Comparison of the GDSRc method with other classifiers on the MIT 67 Scenes dataset.

5.7. Evaluation of the GDSRc method

We present the evaluation of GDSRc when different values of the top T largest dictionary distribution coefficient (v_{ji}) (defined in Section 4.4) are applied for different sizes of the dictionary. The experimental results in figure 4 show that the classification performance improves as the value of T increases regardless of the size of the dictionary. In order to get the best performance, the value of T is usually set as the size of the training data samples for each class.

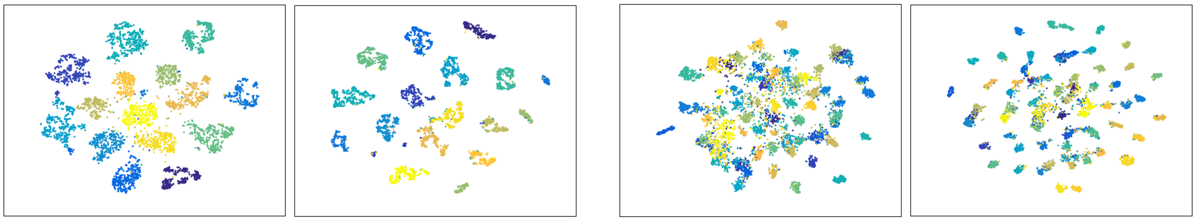
We also present the comparison of the GDSRc method with other classifiers when the same feature representation is learned after applying the GDSR method. Specifically, the k nearest neighbor (KNN) classifier with k set as 3, the linear kernel based SVM and the RBF kernel based SVM are applied for comparison. The experimental results in table 7 show that our proposed GDSRc method achieves better results compared to other classifiers.

6. Conclusion

This paper presents a new generative and discriminative sparse representation (GDSR) method, which leads to a new effective representation and classification schema. In particular, the generative criterion reveals the class conditional probability of each dictionary item and the discriminative criterion applies new within-class and between-class scatter matrices. In addition, a GDSRc classification method is proposed by utilizing both the discriminative and generative information. Experimental results on several classification datasets show the effectiveness of the proposed methods.

Acknowledgements

This work is partially supported by the NSF award 1647170.



(a) Input spatial pyramid features
(15 scenes dataset)

(b) Proposed GDSR features
(15 scenes dataset)

(c) Input CNN features
(MIT-67 scenes dataset)

(d) Proposed GDSR features
(MIT-67 scenes dataset)

Figure 2. The t-SNE visualization of the initial input features and the features extracted after applying the proposed GDSR method.

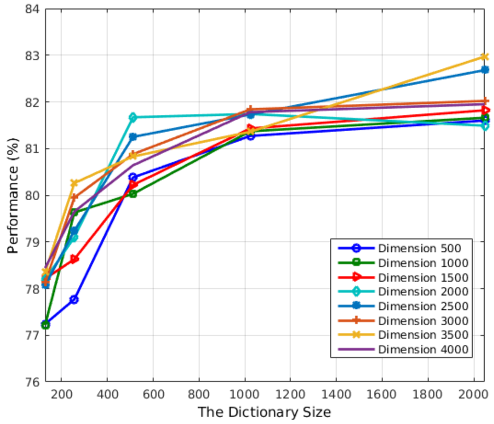


Figure 3. The performance of the proposed GDSR method under different dictionary sizes and feature dimensionality.

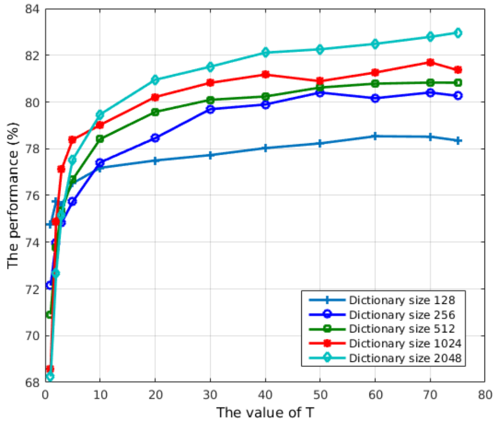


Figure 4. The performance of the proposed GDSRc method when the value of T varies on the MIT-67 indoor scenes dataset.

References

[1] N. Akhtar, F. Shafait, and A. Mian. Discriminative bayesian dictionary learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 2, 6

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009. 4

[3] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 660–667, 2013. 1, 6

[4] S. Gao, I. W.-H. Tsang, and L.-T. Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013. 1, 6

[5] H. Goh, N. Thome, M. Cord, and J.-H. Lim. Learning deep hierarchical visual feature coding. *IEEE Transactions on Neural Networks and Learning Systems*, 2014. 1, 6

[6] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 5, 6

[7] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[8] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493, 1998. 1

[9] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013. 1, 2, 6

[10] I. Jindal, M. Nokleby, and X. Chen. Learning deep networks from noisy labels with dropout regularization. In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 967–972, Dec 2016. 2

[11] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 923–930, 2013. 5

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. 5, 6

[13] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, pages 801–808, 2007. 4, 5

[14] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005. 5, 6

[15] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010. 5

[16] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 851–858, 2013. 5

[17] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[18] J. Liu, C. Gao, D. Meng, and W. Zuo. Two-stream contextualized CNN for fine-grained image classification. In D. Schuurmans and M. P. Wellman, editors, *AAAI 2016*, pages 4232–4233, 2016. 1

[19] L. Liu, C. Shen, L. Wang, A. v. d. Hengel, and C. Wang. Encoding high dimensional local features by sparse coding based fisher vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1143–1151, 2014. 1

[20] Q. Liu and C. Liu. A novel locally linear knn model for visual recognition. In *CVPR*, 2015. 1, 5

[21] Q. Liu and C. Liu. A novel locally linear knn method with application to visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2016. 1, 2, 3

[22] Q. Liu, A. Puthenputhussery, and C. Liu. Novel general knn classifier and general nearest mean classifier for visual classification. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1810–1814, 2015. 1

[23] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 7

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2008. 2

[25] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS*, pages 841–848, 2002. 1

[26] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, pages 1307–1314, 2011. 5

[27] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010. 1, 6

[28] A. Puthenputhussery, Q. Liu, and C. Liu. *Sparse Representation Based Complete Kernel Marginal Fisher Analysis Framework for Computational Art Painting Categorization*, pages 612–627. ECCV 2016, 2016. 1

[29] A. Puthenputhussery, Q. Liu, and C. Liu. A sparse representation model using the complete marginal fisher analysis framework and its applications to visual recognition. *IEEE Transactions on Multimedia*, 19(8):1757–1770, Aug 2017. 1

[30] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009. 5, 7

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 6

[32] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, pages 3400–3407, 2013. 5

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, 2010. 1, 6

[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 1, 3, 6

[35] Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 900–908. 2011. 1

[36] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009. 1, 6

[37] M. Yang, D. Dai, L. Shen, and L. V. Gool. Latent dictionary learning for sparse representation based classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[38] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550, 2011. 6

[39] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, pages 1–24, 2014. 1, 2, 6

[40] S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015. 5, 6

[41] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2698, 2010. 1, 2, 6

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *Arxiv*, 2016. 5, 6

[43] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–495, 2014. 5, 6

[44] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3490–3497, 2012. 2