Vision Recognition using Discriminant Sparse Optimization Learning
Qingxiang Feng, Yicong Zhou, Senior Member, IEEE.

Abstract—To better select the correct training sample and obtain the robust representation of the query sample, this paper proposes a discriminant-based sparse optimization learning model. This learning model integrates discriminant and sparsity together. Based on this model, we then propose a classifier called locality-based discriminant sparse representation (LDSR). Because discriminant can help to increase the difference of samples in different classes and to decrease the difference of samples within the same class, LDSR can obtain better sparse coefficients and constitute a better sparse representation for classification. In order to take advantages of kernel techniques, discriminant and sparsity, we further propose a nonlinear classifier called kernel locality-based discriminant sparse representation (KLDSR). Experiments on several well-known databases prove that the performance of LDSR and KLDSR is better than that of several state-of-the-art methods including deep learning based methods.

Index Terms—Vision Recognition, Image Classification, Object Recognition, Sparse Representation based Classification, Kernel Sparse Representation Classification.

I. INTRODUCTION

With the rapidly development of imaging techniques, image classification and recognition tasks (e.g., face recognition, object recognition, action recognition etc.) attract more and more attention. The main steps of image classification tasks include preprocessing, feature selection and classifiers. In the past few decades, a lot of researchers have worked on these three parts and have made significant improvements. Recently, many researchers pay attention to the second part (feature selection) because deep learning techniques [1], [2] can obtain the quite good feature. However, each technique may have its own bottlenecks. Although deep learning obtains quite good performance, the bottleneck of deep learning will appear sooner or later. We need the help of other techniques (e.g. classifier) to improve the final classification performance. We give an example to explain it.

Example 1: Suppose that there is an examination paper that includes two parts: feature and classifier. The full mark of feature part is 80 points while the full mark of classifier part is 20 points. If an examinee only answers the first part, he/she can’t get the point more than 80.

Example 1 shows that although deep learning is important, other techniques (e.g. classifier) also need to be improved. Generally speaking, classifiers include two categories [3] [4]: parametric-based methods and non-parametric-based methods. The parametric-based methods (e.g., SVM [5]) focus on learning the parameters of a hypothesis classification model from the training data [6]. Then, these methods predict the class labels of unknown data using the learned parametric model. On the contrary, the non-parametric-based methods directly obtain the class labels of unknown data. The well-known non-parametric-based method is the sparse representation classification (SRC) [7]. Compared to other classifiers, SRC and its improved versions obtain better performance for image classification. The sparsity tries to use only the correct training samples (samples have the same label of the testing sample) to constitute a robust representation for classification. However, sparsity doesn’t distinguish correct training samples and incorrect training samples when they have a small difference.

Inspired by linear discriminant analysis (LDA) [8], [9], we know that discriminant can help to increase the difference of samples in different classes and to decrease the difference of samples within the same class. If we increase the discriminant into the optimization learning process of sparsity, we can better select the correct training samples and constitute the robust representation of the query (testing) sample. Based on that, this paper proposes a discriminant-based sparse optimization learning model that integrates the discriminant and sparsity together. Then, we further describe the detail solution procedures of the proposed discriminant-based sparse optimization learning model. Moreover, we propose a classifier called locality-based discriminant sparse representation (LDSR). In order to take advantages of the nonlinear high-dimensional feature, we propose another classifier called kernel locality-based discriminant sparse representation (KLDSR). KLDSR uses the kernel techniques [10] [11] to map the original linear feature to a high-dimensional nonlinear feature for classification. The effectiveness of the proposed classifiers are assessed on several visual classification tasks, which include face recognition on the LFW database [12]; handwritten digit recognition on the MNIST database [13] and USPS dataset [14]; flowers recognition on the Oxford 102 Flowers database [15]; birds recognition on the Caltech-UCSD Birds (CUB200-2011) database [16]; object recognition on the Caltech 256 object databases [17]; and Millions level-based image classification on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset [18].
The main contributions of this paper are as follows:

- Motivated by LDA and SRC, we propose a discriminant-based sparse optimization learning model that integrates the discriminant and sparsity together. The detail solution procedures are also given.
- Using the proposed discriminant-based optimization learning model, we propose a classifier called locality-based discriminant sparse representation (LDSR).
- Motivated by kernel techniques, we further propose kernel locality-based discriminant sparse representation (KLDSR) classifier.
- We evaluate the proposed LDSR and KLDSR on several well-known databases including the million-level database: ImageNet.

II. RELATED WORK

This section reviews the sparse-based methods and linear discriminant analysis (LDA). Several notations are described in Table I.

### A. Sparse-based Methods

Wright et al. proposed sparse representation classification (SRC) in 2009. It represents the testing sample using the linear combination of training samples of all classes. SRC solves $L_1$-minimization optimization problem as

$$||\alpha||_1 \quad s.t. \ X \alpha = x$$

To reduce the computation cost and obtain the better representation, Zhang et al. proposed the collaborative representation based classification (CRC). In CRC, the authors argued that the collaborative representation should be better than the linear combination of training samples. CRC solves the $L_2$-minimization optimization problem as

$$||\alpha||_2 \quad s.t. \ X \alpha = x$$

They are two typical sparse-based methods. A lot improved version of SRC/CRC have been proposed for various visual recognition tasks. They include the probabilistic-based sparse methods [6], discriminant-based dictionary learning [19], [20], kernel-based sparse methods [10] [21], Gabor-based sparse methods [22] and many others [23], [24].

### B. Linear discriminant analysis (LDA)

LDA is a typical machine learning method. It aims at obtaining the largest mean differences between the desired classes. Mathematically, LDA is to maximize the Fisher-ratio criterion [25] as

$$\frac{|O_w|}{|O_b|} \quad O_w \text{ is the within-class scatter matrix, calculated by}$$

$$O_w = \sum_{c=1}^{M} \sum_{i=1}^{N_c} (x^c_i - m_c)(x^c_i - m_c)^T$$

where $m_c$ is the mean sample of the $c^{th}$ class. $O_b$ is between-class scatter matrix, computed by

$$O_b = \sum_{c=1}^{M} (m_c - m)(m_c - m)^T$$

where $m$ is the mean sample of all classes.

III. DISCRIMINANT-BASED SPARSE OPTIMIZATION LEARNING MODEL

This section proposes a discriminant-based sparse optimization learning model. To better explain this learning model, we divide this section into three subsections. Section III-A gives the motivation. Section III-B proposes a discriminant-based sparse optimization learning model. Section III-C introduces the detail solution procedures of this optimization learning model.

#### A. Motivation

The sparsity tries to select the correct training samples (samples have the same label of the testing sample) and to constitute a robust representation of the testing sample. However, variations (e.g. lights, views, occlusions, pose, background etc.) in the collected images make it challenging to obtain a robust representation model for image classification. Inspired by LDA, we know that discriminant can help to increase the difference of samples in different classes and to decrease the difference of samples within the same class. If we increase the discriminant into the optimization process of sparsity, the correct training samples and the incorrect training samples would have a larger difference. Then, we can better select the correct training samples to constitute the robust representation of the query sample. Fig. 1 gives an example to explain the motivation. Observing Fig. 1, we know that if we increase the discriminant in the optimization learning process of sparsity, the discriminant-sparse-based coefficients in the same class would have a smaller difference, and the discriminant-sparse-based coefficients in different classes would have a bigger difference. The next section will describe how to increase the discriminant in the optimization learning process of sparsity.

#### B. Discriminant-based Sparse Learning Model

Based on the sparse representation and LDA, we propose a discriminant-based sparse learning model, its objective function is defined as

$$\min_{\alpha} ||x - X\alpha||_2^2 + \lambda ||\alpha||_2^2 + \eta S_w + \gamma S_h$$

where $S_w$ means the difference of the samples within the same class. $S_h$ denotes the difference of the samples in different classes.

**TABLE I**

| Notation | Explanation |
|----------|-------------|
| $X$      | Entire training set |
| $x^c$    | The $i^{th}$ sample of the $c^{th}$ class |
| $w$      | Dimension of a sample |
| $q$      | Number of samples of the $c^{th}$ class |
| $M$      | Number of classes |
| $x$      | Testing sample |
is the corresponding sparse coefficient of $X$ where $X$ as $S$.

Motivated by [26], we know that minimizing $||x_i - x_j||_2^2$ in different classes. That is, we want to minimize $||x_i - x_j||_2^2$ in different classes, we try to minimize the correlation of different samples. To minimize the difference of the samples within the same class, we try to minimize the distance between each sample and the corresponding class’s sparse representation. Specifically, $S$ is the sample of the $i$th class, $\alpha_c \in R^{N_x \times 1}$ is the corresponding sparse coefficient of $X_c$, and $\alpha_c^i \in R^{1 \times 1}$ is the $i$th element of $\alpha_c$.

1) How to obtain $S_w$: In SRC, each class has a sparse representation. To minimize the difference of the samples within the same class, we try to minimize the distance between each sample and the corresponding class’s sparse representation. Specifically, $S_w$ can be computed by

$$S_w = \sum_{c=1}^{N_c} \sum_{i=1}^{N_x} (x_i^c \alpha_i^c - X_c \alpha_c)^T (x_i^c \alpha_i^c - X_c \alpha_c)$$

$$= \sum_{c=1}^{N_c} \sum_{i=1}^{N_x} ||x_i^c \alpha_i^c - X_c \alpha_c||^2$$

where $X_c$ denotes the samples of the $c$th class, $\alpha_c \in R^{N_x \times 1}$ is the corresponding sparse coefficient of $X_c$, and $\alpha_c^i \in R^{1 \times 1}$ is the $i$th element of $\alpha_c$.

2) How to obtain $S_b$: To increase the difference of samples in different classes, we try to minimize the correlation of different classes. That is, we want to minimize $(X_i \alpha_i)^T (X_j \alpha_j)$. Motivated by [26], we know that minimizing $||X_i \alpha_i||^2$ is also good for sparsity. Thus, $S_b$ can be obtained by

$$S_b = \sum_{i=1}^{M} \sum_{j=1}^{M} (||X_i \alpha_i||^2 + 2(X_i \alpha_i)^T (X_j \alpha_j) + ||X_j \alpha_j||^2)$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{M} ||X_i \alpha_i + X_j \alpha_j||^2$$

Consider $S_w$ and $S_b$, the above objective can be rewritten as

$$\min_{\alpha} ||x - X\alpha||^2 + \lambda ||\alpha||^2 + \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_x} ||x_i^c \alpha_i^c - X_c \alpha_c||^2$$

$$+ \eta \sum_{i=1}^{M} \sum_{j=1}^{M} ||X_i \alpha_i + X_j \alpha_j||^2$$

The objective function in Eq. (8) is differentiable. Thus, the optimal solution of Eq. (8) can be treated as the stationary point of the objective function. Let $G(\alpha) = \min_{\alpha} ||x - X\alpha||^2 + \lambda ||\alpha||^2 + \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_x} ||x_i^c \alpha_i^c - X_c \alpha_c||^2 + \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} ||X_i \alpha_i + X_j \alpha_j||^2$. The derivative of the first term of $G(\alpha)$ is

$$\frac{\partial(||x - X\alpha||^2)}{\partial(\alpha)} = -2X^T (x - X\alpha)$$

The derivative of the second term of $G(\alpha)$ is

$$\frac{\partial(\lambda ||\alpha||^2)}{\partial(\alpha)} = 2\lambda \alpha$$

The derivative of the third term of $G(\alpha)$ is complex. Let $f(\alpha) = \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_x} ||x_i^c \alpha_i^c - X_c \alpha_c||^2$, we rewrite the $f(\alpha)$ as

$$f(\alpha) = \eta \sum_{i=1}^{M} \sum_{j=1}^{M} ||X_i \alpha_i + X_j \alpha_j||^2$$

C. Solution Procedures of Eq. (8)

Fig. 1. The main idea of the proposed method. Discriminant can help to increase the difference of samples in different classes and decrease the difference of samples in the same class. Given a testing sample and several training samples. If we increase the discriminant in the optimization process of sparsity, the discriminant-sparse-based coefficients in the same class would have smaller difference, and the discriminant-sparse-based coefficients in different class would have bigger difference, compared to sparse-based coefficients.
Using the rewritten \( f(\alpha) \) obtain \( \partial f(\alpha) \), solve the problem, we firstly compute the partial derivative. To directly. To obtain \( \partial f(\alpha)/\partial(\alpha_c) \) as

\[
\frac{\partial f(\alpha)}{\partial(\alpha_c)} = \gamma \left( \sum_{i=1}^{M} \sum_{j \neq c} \|X_i \alpha_i + X_j \alpha_j\|^2 \right) + 2 \eta \left( \sum_{i=1}^{N_1} \|X_i^T X_{1,i}\| \right) + 4 \gamma X^T X \alpha + 4 \gamma (M - 2) X_{\alpha,c} + \sum_{i=1}^{N_1} X_i \alpha_i + X_0 \alpha + X_0 \alpha)
\]

Using Eqs. (9), (10), (12) and (15), we can obtain the \( \partial G(\alpha)/\partial(\alpha) \) as

\[
\frac{\partial G(\alpha)}{\partial(\alpha)} = -2 X^T (x - X \alpha) + 2 \lambda \alpha + 4 \gamma X^T X \alpha
\]

Moreover, the derivative of the fourth term of \( g(\alpha) \) is complex. Because \( f(\alpha) = \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} \|X_i \alpha_i + X_j \alpha_j\|^2 \) doesn’t explicitly include \( \alpha \), we can’t compute \( \partial f(\alpha)/\partial(\alpha) \) directly. To solve the problem, we firstly compute the partial derivative \( \partial f(\alpha)/\partial(\alpha_c) \), and then use all \( \partial f(\alpha)/\partial(\alpha_c) \) to obtain \( \partial f(\alpha)/\partial(\alpha) \). To compute the derivative \( \partial f(\alpha)/\partial(\alpha_c) \), we rewrite the \( f(\alpha) = \gamma \sum_{i=1}^{M} \sum_{j \neq c} \|X_i \alpha_i + X_j \alpha_j\|^2 \) as follows

\[
f(\alpha) = \gamma \sum_{i=1}^{M} \sum_{j \neq c} \|X_i \alpha_i + X_j \alpha_j\|^2
\]

where \( X_{c,i} = \begin{bmatrix} x_i^c & \ldots & 0 & \ldots & x_i^c \end{bmatrix} \in \mathbb{R}^{q \times N_c}, c = 1, 2, \ldots, M, i = 1, 2, \ldots, N_c \).
Algorithm 1 Locality-based Discriminant Sparse Representation

Inputs: The entire training set $X$ with $M$ classes $X_c \in R^{p \times L_c}$ for $c = 1, 2, ..., M$ and a testing sample $x \in R^{p \times 1}$.

Output: class label of $x$.

1: Use $X$ and $x$ to solve discriminant-based sparse coefficient $\alpha$ by

$$
\alpha = \left( X^T X + \lambda I + \gamma \left( \begin{array}{ccc}
X_1^T X_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & X_M^T X_M
\end{array} \right) \right)^+ \left( \begin{array}{c}
\sum_{i=1}^{N_c} X_{1,i}^T X_{1,i} \\
\vdots \\
\sum_{i=1}^{N_c} X_{M,i}^T X_{M,i}
\end{array} \right) x
$$

where $X_{c,i} = [ x_{c,i} \ldots x_{c,i-1} 0 x_{c,i+1} \ldots x_{c,N_c} ] \in R^{q \times N_c}$, $c = 1, 2, \ldots M$, $i = 1, 2, \ldots N_c$.

2: Compute the sparse-based distance between the testing sample and training samples by $d_i^c = ||x - x_i^c \alpha_i^c||$. Constitute the locality-based training set $Y$ by selecting $s$ samples from $X$.

3: Use $Y$ and $x$ to solve discriminant-based sparse coefficient $\beta$ by

$$
\beta = \left( Y^T Y + \lambda I + \gamma \left( \begin{array}{ccc}
Y_1^T Y_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & Y_M^T Y_M
\end{array} \right) \right)^+ \left( \begin{array}{c}
\sum_{i=1}^{N_c} Y_{1,i}^T Y_{1,i} \\
\vdots \\
\sum_{i=1}^{N_c} Y_{M,i}^T Y_{M,i}
\end{array} \right) x
$$

where $Y_{c,i} = [ y_{c,i} \ldots y_{c,i-1} 0 y_{c,i+1} \ldots y_{c,N_c} ] \in R^{q \times L_c}$, $c = 1, 2, \ldots M$, $i = 1, 2, \ldots L_c$, $L_c$ is number of samples of $Y_c$.

4: The distance between $x$ and the $c^{th}$ class is computed by $s_c = \frac{||x - Y_c \beta_c||}{||\beta||}$. Classify the testing sample $x$ into the class with the minimization distance by $c* = \arg \min(s_c)$.

The solution can be obtained when the condition $\frac{\partial G(\alpha)}{\partial (\alpha)} = 0$ is satisfied. That is

$$
0 = -2X^T(x - X\alpha) + 2\lambda \alpha + 4\gamma X^T X \alpha + 2\eta \left( \sum_{i=1}^{N_c} X_{1,i}^T X_{1,i} \right) \alpha + 2\eta \left( \begin{array}{ccc}
\sum_{i=1}^{N_c} X_{1,i}^T X_{1,i} \\
\vdots \\
\sum_{i=1}^{N_c} X_{M,i}^T X_{M,i}
\end{array} \right) \alpha + 4\gamma(M - 2) \left( \begin{array}{ccc}
X_1^T X_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & X_M^T X_M
\end{array} \right) \alpha
$$

Solving the Eq. (17), we can obtain the optimal solution as

$$
\alpha = \left( X^T X + \lambda I + \eta H^I + 2\gamma(M - 2)H^{II} \right) x
$$

where $H^I$ is

$$
H^I = \left( \begin{array}{ccc}
\sum_{i=1}^{N_c} X_{1,i}^T X_{1,i} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sum_{i=1}^{N_c} X_{M,i}^T X_{M,i}
\end{array} \right)
$$

and $X_{c,i} = [ x_{c,i} \ldots x_{c,i-1} 0 x_{c,i+1} \ldots x_{c,N_c} ] \in R^{q \times N_c}$, $c = 1, 2, \ldots M$, $i = 1, 2, \ldots N_c$. $H^{II}$ is

$$
H^{II} = \left( \begin{array}{ccc}
X_1^T X_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & X_M^T X_M
\end{array} \right)
$$

IV. LOCALITY-BASED DISCRIMINANT SPARSE REPRESENTATION

Using the discriminant-based sparse optimization learning model in Section III, this section proposes a classifier, called locality-based discriminant sparse representation (LDSR). Sparse representation tries to obtain a representation constituted only by the training samples with the same label of the testing sample. However, the sparsity constraint cannot guarantee it because the samples of other classes often have the effect of constituting the representation. To solve this, we try to reserve the training samples with the same label of the testing sample, and to decrease the training samples with the different label of the testing sample. Motivated by this, we try to obtain the locality-based training set as follows. Constitute the discriminant-based optimization learning model by Eq. (8). Compute the sparse-based distance between the testing sample and training samples by

$$
d_i^c = ||x - x_i^c \alpha_i^c||
$$

Select $s$ samples with the smallest distances to constitute the locality-based training set $Y$. Using $Y$, we constitute the discriminant-based optimization function as

$$
\min_{\beta} ||x - Y \beta||^2_2 + \lambda ||\beta||^2_2 + \eta \sum_{c=1}^{M} \sum_{i=1}^{L_c} ||y_i^c \beta_i^c - Y_c \beta_c||^2_2 + \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} ||Y_i \alpha_i + Y_j \alpha_j||^2
$$

where $Y_c$ means the samples in the $c^{th}$ class, $L_c$ is the number of samples of $Y_c$, $y_i^c$ is the $i^{th}$ sample of the $c^{th}$ class, $i = 1, 2, \ldots L_c$.

Solve Eq. (22) and obtain the sparse coefficient as

$$
\beta = (Y^TY + \lambda I + \eta P^I + 2\gamma(M - 2)P^{II}) Y^T x
$$
where $P^I$ is

$$P^I = \begin{pmatrix}
\sum_{i=1}^{N_c} Y_{1,i}^T Y_{1,i} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sum_{i=1}^{N_c} Y_{M,i}^T Y_{M,i}
\end{pmatrix}$$

(24)

and $Y_{c,i} = [x_{i1}^c \ldots x_{i_{c-1}}^c \quad 0 \quad x_{i_{c+1}}^c \ldots x_{i_{N_c}}^c] \in \mathbb{R}^{q \times N_c}$, $c = 1, 2, \ldots M$, $i = 1, 2, \ldots N_c$, $P^{II}$ is

$$P^{II} = \begin{pmatrix}
Y_{1,i}^T Y_{1,i} & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & Y_{M,i}^T Y_{M,i}
\end{pmatrix}$$

(25)

**Compute the distance between x and the c^th class as**

$$s_c = \frac{||x - Y_c \beta_c||}{|| \beta_c ||}$$

(26)

**Classify the testing sample x with the minimization distance by**

$$c = \arg \min (s_c)$$

(27)

The detail procedures of LDSR are summarized in Algorithm 1.

**V. PROPOSED KLDSR**

In this Section, we propose the kernel locality-based Discriminative Sparse Representation (KLDSR) classifier. The detail processes of KLDSR are summarized in Algorithm 2.

**A. Kernel trick**

In order to map a linear feature to a high-dimensional nonlinear feature, the kernel techniques have been proposed. In this section, we utilize the most popular Gaussian radial basis function (RBF) kernel for classification. The RBF kernel can be described as

$$k(x, y) = \phi(x)^T \phi(y) = \exp\left(-\frac{||x - y||^2}{\sigma}\right)$$

(28)

where $x$ and $y$ denote any two original samples, $\sigma$ denotes a parameter. In kernel methods, $\phi(*)$ is unknown. We can use only $k(*, *)$ to access the feature space.

**B. Nonlinear objective function**

We suppose that there is a nonlinear feature mapping function $\Phi(.) : \mathbb{R}^q \rightarrow \mathbb{R}^Q (q << Q)$. This function maps the testing sample $x$ and training set $X$ into a high-dimensional feature space as

$$x \rightarrow \Phi(x)$$

$$X \rightarrow \Phi(X) = [\Phi(x_1) \ldots \Phi(x_i) \ldots \Phi(x_L)]$$

(29)

The nonlinear objective function can be described as

$$\min_{\alpha} \|\phi(x) - \phi(X)\alpha\|_2^2 + \lambda \|\alpha\|_2^2$$

$$+ \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_c} ||\phi(x_i^c)\alpha_i^c - \phi(x_i)\alpha_c||_2^2$$

$$+ \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} ||\phi(X_i)\alpha_i + \phi(X_j)\alpha_j||_2^2$$

(30)

Because the dimension of nonlinear information is quite high, it is difficult to solve the above formula. We use the kernel trick and obtain the following objective function.

$$\min_{\alpha} \|\phi(X)^T \phi(x) - \phi(X)^T \phi(x)\alpha\|_2^2 + \lambda \|\alpha\|_2^2$$

$$+ \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_c} ||\phi(X)^T \phi(x_i^c)\alpha_i^c - \phi(X)^T \phi(x_i)\alpha_c||_2^2$$

$$+ \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} ||\phi(X)^T \phi(X_i)\alpha_i + \phi(X)^T \phi(X_j)\alpha_j||_2^2$$

(31)

Using the kernel technique, we get a kernel matrix $K = \Phi(X)^T \Phi(X)$ as

$$K = \begin{bmatrix}
k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_L) \\
k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_L) \\
\vdots & \vdots & \ddots & \vdots \\
k(x_L, x_1) & k(x_L, x_2) & \ldots & k(x_L, x_L)
\end{bmatrix}$$

(32)

and $K_c = \Phi(X)^T \Phi(x_i^c)$ is the corresponding columns $(X_c)$ of $K$, and $K_c^i = \Phi(X)^T \Phi(x_i^c)$

$$K_c^i = [\ k(x_1, x_1^c) \ k(x_2, x_2^c) \ldots \ k(x_L, x_L^c) \ ]^T$$

(33)

and a testing vector $k(., x) = \Phi(Y)^T \Phi(x)$ as

$$k(., x) = \phi(X)^T \phi(x)$$

$$= [ \ k(x_1, x) \ k(x_2, x) \ldots \ k(x_L, x) \ ]^T$$

(34)

Consider Eqs. 32-34, the nonlinear objective function can be rewritten as

$$\min_{\alpha} \|k(.,x) - K\alpha\|_2^2 + \lambda \|\alpha\|_2^2$$

$$+ \eta \sum_{c=1}^{N_c} \sum_{i=1}^{N_c} ||K_c^i\alpha_i^c - K_c\alpha_c||_2^2$$

$$+ \gamma \sum_{i=1}^{M} \sum_{j=1}^{M} ||K_i\alpha_i + K_j\alpha_j||_2^2$$

(35)

Solving Eq. 35, we can obtain the optimal solution as

$$\alpha = (K^T K + \lambda I + \eta A^I + 2\gamma(M - 2)A^{II}) K^T x$$

(36)

where $A^I$ is

$$A^I = \begin{pmatrix}
N_c & \cdots & 0 \\
0 & \cdots & M \\
0 & \cdots & N_c \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{pmatrix}$$

(37)

and $K_{c,i} = \phi(X)^T [\ x_i^c \ldots x_i^{c-1} \quad 0 \quad x_i^{c+1} \ldots x_i^{N_c} \ ]$, $c = 1, 2, \ldots M$, $i = 1, 2, \ldots N_c$, and $A^{II}$ is

$$A^{II} = \begin{pmatrix}
K_c^1 & \cdots & 0 \\
0 & \cdots & K_c^M \\
0 & \cdots & K_c^M
\end{pmatrix}$$

(38)

Compute the sparse-based distance between the testing sample and training samples by

$$d_c^i = ||\phi(x) - \phi(x_i^c)||_2$$

$$= \sqrt{(\phi(x) - \phi(x_i^c))^T (\phi(x) - \phi(x_i^c))}$$

$$= \sqrt{k(x, x) - 2k(x, x_i^c)\alpha + \alpha^T k(x_i^c, x_i^c)\alpha}$$

(39)
**Algorithm 2 Kernel Locality-based Discriminant Sparse Representation**

Inputs: The entire training set $X$ with $M$ classes $X_c \in \mathbb{R}^{N_c \times d}$ for $c = 1, 2, ..., M$ and a testing sample $x \in \mathbb{R}^d$.

Output: class label of $x$.

1. Constitute the kernel matrix $K$ and testing kernel vector $k(., x)$ by

$$K = \begin{bmatrix}
  k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_{N_c}) \\
  k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_{N_c}) \\
  \vdots & \vdots & \ddots & \vdots \\
  k(x_{N_c}, x_1) & k(x_{N_c}, x_2) & \cdots & k(x_{N_c}, x_{N_c})
\end{bmatrix}$$

$$k(., x) = [k(x_1, x) k(x_2, x) \cdots k(x_{N_c}, x)]^T$$

2. Utilize $K$ and $k(., x)$ to solve discriminant-based sparse coefficient $\alpha$ by

$$\alpha = \left( K^T K + \lambda I + \gamma \begin{bmatrix} K_1^T K_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_M^T K_M \end{bmatrix} \right) + \eta \left( \begin{bmatrix} \sum_{i=1}^{N_c} U_{1,i}^T U_{1,i} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{i=1}^{N_c} U_{M,i}^T U_{M,i} \end{bmatrix} \right) K^T k(., x)$$

where $K_{c,i} = \phi(X)^T [x_i^2 \ x_{i-1}^2 \ 0 \ x_{i+1}^2 \ \cdots \ x_{N_c}^2]$, $c = 1, 2, \ldots, M$, $i = 1, 2, \ldots, N_c$.

3. Compute the nonlinear sparse-based distance between the testing sample and training samples by

$$d_i^u = ||\phi(x) - \phi(x_i)\alpha|| = \sqrt{\phi(x)^T \phi(x_i) \alpha^T \phi(x) - \phi(x_i)^T \phi(x) \alpha} = \sqrt{k(x, x_i^2) \alpha^T \alpha - 2k(x, x_i^2) \alpha}$$

Select $s$ samples as $Y$ and constitute a locality-based kernel matrix $U$ and a testing kernel vector $u(., x)$. The calculation methods of $U$ and $u(., x)$ are similar to those of $K$ and $k(., x)$.

4. Use $U$ and $u(., x)$ to solve discriminant-based sparse coefficient $\beta$ by

$$\beta = \left( U^T U + \lambda I + \gamma \begin{bmatrix} U_1^T U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_M^T U_M \end{bmatrix} \right) + \eta \left( \begin{bmatrix} \sum_{i=1}^{N_c} U_{1,i}^T U_{1,i} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{i=1}^{N_c} U_{M,i}^T U_{M,i} \end{bmatrix} \right) U^T u(., x)$$

where $U_{c,i} = \phi(Y)^T [y_i^c \ y_{i-1}^c \ 0 \ y_{i+1}^c \ \cdots \ y_{L_c}^c]$, $c = 1, 2, \ldots, M$, $i = 1, 2, \ldots, L_c$, $L_c$ is number of samples of $Y_c$.

5. The distance between $x$ and the $c^{th}$ class is computed by $s_c = \frac{||u(., x) - U_c \beta_c||}{||\beta_c||}$. Classify the testing sample $x$ into the class with the minimization distance by

$$c^* = \operatorname{arg\,min}_c(s_c)$$

Select $s$ samples as $Y$ and constitute a locality-based kernel matrix $U$ and a testing kernel vector $u(., x)$ and $k(., x)$. The calculation methods of $U$ and $u(., x)$ are similar to those of $K$ and $k(., x)$. Using $U$ and $u(., x)$, the nonlinear-discriminant-based optimization function is described as

$$\min_\beta ||u(., x) - U \beta||^2 + \lambda ||\beta||^2$$

$$+ \eta \sum_{i=1}^{M} \sum_{k=1}^{K_c} ||U_c^k \beta_c^k - U_c \beta_c||^2$$

$$+ \gamma \sum_{i=1}^{N_c} \sum_{j=1}^{M} ||U_{c,i} \alpha_i + U_{c,j} \alpha_j||^2$$

where $U_c$ indicates a sub-kernel-matrix constituted by samples in the $c^{th}$ class, $L_c$ is the number of samples of $U_c$.

Solving Eq. (40), we can obtain the optimal solution as

$$\alpha = (U^T U + \lambda I + \eta B^T + 2\gamma(M - 2)B^{II})^T x$$

where $B^I$ is

$$\begin{bmatrix}
  \sum_{i=1}^{N_c} U_{1,i}^T U_{1,i} & \cdots & 0 \\
  \vdots & \ddots & \vdots \\
  0 & \cdots & \sum_{i=1}^{N_c} U_{M,i}^T U_{M,i}
\end{bmatrix}$$

and $U_{c,i} = \phi(Y)^T [y_i^c \ y_{i-1}^c \ 0 \ y_{i+1}^c \ \cdots \ y_{L_c}^c]$, $c = 1, 2, \ldots, M$, $i = 1, 2, \ldots, L_c$, and $B^{II}$ is

$$B^{II} = \begin{bmatrix}
  U_1^T U_1 & \cdots & 0 \\
  \vdots & \ddots & \vdots \\
  0 & \cdots & U_M^T U_M
\end{bmatrix}$$

Compute the distance between $x$ and the $c^{th}$ class as

$$s_c = \frac{||u(., x) - U_c \beta_c||}{||\beta_c||}$$

Classify the minimization distance by

$$c^* = \operatorname{arg\,min}_c(s_c)$$

The detail procedures of KLDSR are summarized in Algorithm 2.

VI. EXPERIMENTAL RESULTS

The performance of the proposed LDSR and KLDSR classifiers is evaluated on several vision recognition databases: face recognition on the LFW database [12], handwritten digit recognition on the MNIST database [13] and USPS dataset [14], flowers recognition on the Oxford 102 Flowers database
### TABLE II

**Recognition rates (RR) of several classifiers on LFW face database**

| Classifier | Accuracy (%) |
|------------|--------------|
| SVM        | 43.30        |
| NSC        | 43.80        |
| SRC        | 44.10        |
| KSRC       | 44.80        |
| CRC        | 44.30        |
| FDDL       | 42.00        |
| MCT        | 44.90        |
| NDSR       | 45.70        |
| ProCRC     | 45.80        |
| LDSR       | 46.40        |
| KLDSR      | 46.80        |

### TABLE III

**Recognition rates (RR) of several classifiers on the MNIST database**

| Classifier | 50   | 100  | 300  |
|------------|------|------|------|
| SVM        | 89.35| 92.10| 94.88|
| NSC        | 91.06| 92.86| 85.29|
| CRC        | 72.21| 82.22| 86.54|
| SRC        | 80.12| 85.63| 89.30|
| KSRC       | 80.32| 85.86| 89.88|
| CROC       | 91.06| 92.86| 89.93|
| NDSR       | 91.64| 94.06| 95.18|
| ProCRC     | 91.84| 94.00| 95.48|
| LDSR       | 92.62| 94.73| 96.03|
| KLDSR      | 92.91| 94.89| 96.10|

### TABLE IV

**Recognition rates (RR) of several classifiers on the USPS database**

| Classifier | 50   | 100  | 300  |
|------------|------|------|------|
| SVM        | 93.46| 95.31| 96.30|
| NSC        | 93.48| 93.25| 87.85|
| CRC        | 89.89| 91.67| 92.79|
| SRC        | 92.58| 93.99| 95.86|
| KSRC       | 92.48| 93.78| 95.96|
| CROC       | 93.48| 93.25| 91.87|
| NDSR       | 93.68| 95.31| 96.25|
| ProCRC     | 93.84| 95.62| 96.43|
| LDSR       | 94.26| 96.12| 96.87|
| KLDSR      | 94.48| 96.31| 96.91|

---

**A. Face recognition**

LFW face database is used in this experiment. Following the protocol in [27], we apply 158 subjects that have no less than ten samples for evaluation. The experiment here is set as follows: For each person, 5 samples are randomly selected to form the training set, while other 2 samples are exploited for testing. The SRC [7], SVM [5], FDDL [19], MCT [27], NSC [28], ProCRC [6] and CRC [29] algorithms are chosen for comparison. Table II illustrates the comparison results of all methods. Compared to these existing methods, KLDSR has more than 1% improvement.

**B. Handwritten Digit Recognition**

This section uses two databases (MNIST dataset and USPS dataset) to evaluate the performance of proposed DSR for handwritten digit recognition. These two databases are described as follows.

**MNIST dataset:** This handwritten digit dataset has 70000 images: 60,000 handwritten digit images are the training set and the rest 10,000 handwritten digit images are testing set. There are 10 digits (0-9) in total. Each handwritten digit image is with the size of 28 × 28.

**USPS dataset:** This handwritten digit dataset has 9306 images, the training set includes 7291 images and the rest 2007 handwritten digit images are testing set. It also has ten digits (0-9). The size of each image is 16 × 16.

This experiment follows the setting in ref. [6]: We randomly select 50, 100, and 300 samples from each digit (class), and use all samples of the testing set for testing. Tables III and IV list the experiment results. We can observe that the proposed LDSR and KLDSR classifiers outperform the other comparison methods. Moreover, the recognition rates of LDSR and KLDSR increase consistently with the increasing number of training image samples. However, NSC has the dropping recognition rate in the same situation.

**C. Performance on deep learning feature**

This section assesses the proposed classifiers with the deep learning feature on three challenging databases, namely the Oxford 102 Flowers database, Caltech-UCSD Birds (CUB200-2011) database and Caltech-256 database. In the following experiment, VGG-verydeep-19 is employed to extract CNN feature (namely VGG19 features) [33]. The activations of the penultimate layer is used as local features. They are extracted from 5 scales $2^s, s = -1, -0.5, 0, 0.5, 1$. We pool all local features together regardless of scales and locations. The final feature dimension of each sample in these databases is 4,096.

1) **Flowers Recognition:** Oxford 102 Flowers database has 8189 flower images of 102 categories. Different flowers have different scales, pose and lighting conditions. Due to large variations within the category and small difference of different categories, this database is quite challenging and is also widely-used for fine-grained image recognition. We follow the settings in [6]. Table V describes the experiment results. The left part of Table V gives the experiment results of the...
### Table V: Recognition rates (RR) of several methods on the Oxford 102 Flowers database with VGG19-based deep feature

| Classifier | RR | State-of-the-art Methods | RR |
|------------|----|--------------------------|----|
| Softmax    | 87.3 | NAC [2]                  | 95.3 |
| SVM        | 90.9 | OverFeat [35]            | 86.8 |
| Kernel SVM | 92.2 | GMP [36]                 | 84.6 |
| NSC        | 90.1 | DAS [37]                 | 80.7 |
| CRC        | 93.0 | BigCos set [38]          | 79.4 |
| SRC        | 93.2 |                          |     |
| KSRC       | 93.3 |                          |     |
| CROC       | 93.1 |                          |     |
| ProCRC     | 94.8 |                          |     |
| LDSR       | 95.3 |                          |     |
| KLDSR      | 95.7 |                          |     |

The proposed LDSR, KLDSR and several well-known classifiers; the right part of Table [V] lists the results of several state-of-the-art methods for flowers recognition. From Table [V], we can know that KLDSR obtains the highest recognition rate with the VGG19 features among compared classifiers and the state-of-the-art methods.

2) Birds Recognition: The Caltech-UCSD Birds (CUB200-2011) database has 11,788 images of 200 bird species. Because bird species are high degree of similarity, this database is quite challenging and is widely-used for fine-grained image recognition. We follow the settings in [10]. Around 30 samples of each bird species are used as training set, the resting images are used as testing set. The experiment results are shown in Table [VI]. The left part of Table [VI] lists the experiment results of the proposed LDSR and KLDSR using several well-known classifiers; the right part of Table [VI] lists the experiment results of several state-of-the-art methods for birds recognition. From Table [VI] we can see that KLDSR obtains the highest recognition rate with the VGG19 features among compared classifiers. Moreover, KLDSR obtains the second highest among state-of-the-art methods.

3) Object Recognition: The Caltech-256 dataset has 256 object classes. Each object class contains at least 80 images. This database has 30,608 images in total. To evaluate the performance of LDSR and KLDSR for object recognition, we follow Ref. [6], randomly select 30 images for training, the rest images are used for testing. Table [VII] lists the experiment results. All comparison methods are classifiers. As we can observe, LDSR and KLDSR have at least 1.0% improvements compared to the third-best method (ProCRC). Afterwards, we compare the proposed LDSR and KLDSR with the state-of-the-art methods including four deep learning-based methods.

### Table VI: Recognition rates (RR) of several methods on the Caltech-UCSD Birds (CUB200-2011) database with VGG19-based deep feature

| Classifier | RR | State-of-the-art Methods | RR |
|------------|----|--------------------------|----|
| Softmax    | 72.1 | NAC [2]                  | 81.0 |
| SVM        | 75.4 | PN-CNN [29]              | 75.7 |
| Kernel SVM | 76.6 | FV-CNN [40]              | 66.7 |
| NSC        | 74.5 | POOF [41]                | 56.9 |
| CRC        | 76.2 |                          |     |
| SRC        | 76.0 |                          |     |
| KSRC       | 76.2 |                          |     |
| CROC       | 76.2 |                          |     |
| ProCRC     | 78.3 |                          |     |
| LDSR       | 79.2 |                          |     |
| KLDSR      | 79.5 |                          |     |

### Table VII: Recognition rates (RR) of several classifiers on the and Caltech 256 object database with VGG19-based deep feature

| Classifier | RR |
|------------|----|
| Softmax    | 75.5 |
| SVM        | 80.1 |
| Kernel SVM | 81.3 |
| NSC        | 80.2 |
| CRC        | 81.1 |
| SRC        | 81.3 |
| KSRC       | 81.7 |
| CROC       | 81.7 |
| ProCRC     | 83.3 |
| LDSR       | 84.3 |
| KLDSR      | 84.7 |

The results, we see that the proposed classifiers have at least 1% improvement compared to all state-of-the-art methods.

### D. Performance on ImageNet

The [6] may have the potential scalability problem when the quite large scale dataset is used [6], such as ImageNet database. They may not be feasible to load millions of samples into memory and may not be easy to solve the inverse of a matrix with millions-dimension. To address this problem, the dictionary learning (DL) techniques are used. We follow [6], one simple DL model is used, which is described as

$$
\min_{D_k, A_k} \|X_k - D_k A_k\|_F^2 + \tau \|A_k\|_F^2
$$

where $\tau$ is a constant. Using the above equation, the dictionary $D_k$ will replace the original training set $X_k$. The dimension of samples of $D_k$ is the same as that of $X_k$. However, the number of samples of $D_k$ is different from that of $X_k$. For example, ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset contains 1.2M+ training images with 1000 classes (around 1300 images per class). Using the above
TABLE VIII
Recognition rates (RR) of several state-of-the-art methods on the and Caltech 256 object database with VGG19-based deep feature

| Classifier | 15    | 30    | 45    | 60    |
|------------|-------|-------|-------|-------|
| MHMP [42]  | 40.50 | 48.00 | 51.90 | 55.20 |
| IFV [43]   | 34.70 | 40.80 | 45.00 | 47.90 |
| LLC [44]   | 34.36 | 41.19 | 45.31 | 47.68 |
| ScSPM [45] | 27.73 | 34.02 | 37.46 | 40.14 |
| ZF [46]    | 65.70 | 70.60 | 72.70 | 74.20 |
| LLNMC [47] | 68.32 | 71.89 | 74.13 | 75.47 |
| LLKNC [47] | 68.55 | 72.09 | 74.07 | 75.36 |
| CNN-S [48] | -     | -     | -     | 77.61 |
| VGG19 [33] | -     | -     | -     | 85.10 |
| NAC [4]    | -     | -     | -     | 84.10 |
| ProCRC [6] | 80.20 | 83.30 | 84.90 | 86.10 |
| LDSR       | 81.20 | 84.19 | 86.00 | 87.14 |
| KLDSR      | 81.36 | 84.50 | 86.13 | 87.36 |

TABLE IX
Recognition rates (RR) of several classifiers on ImageNet

| Classifier | Top 5 | Top 1 |
|------------|-------|-------|
| Softmax [34] | 80.4 | 57.4 |
| SVM [5]     | 79.7 | 55.8 |
| NSC [28]    | 77.4 | 53.2 |
| CRC [29]    | 78.5 | 54.3 |
| SRC [7]     | 78.7 | 54.1 |
| KSR [10]    | 78.8 | 54.3 |
| CROC [30]   | 78.8 | 54.4 |
| ProCRC [6]  | 80.1 | 56.3 |
| LDSR        | 80.7 | 56.8 |
| KLDSR       | 80.9 | 57.1 |

The equation, the number of samples of each class is reduced to 50. That is, we have only 50K training images and 50K testing images.

In this experiment, the ImageNet database is used. We follow Ref. [5] and use AlexNet feature extracted by Caffe [1] (feature dimension is 4096). Table IX shows the recognition rates of several classifiers with top-1 and top-5 schemes. From Table IX, we know that KLDSR obtains the best performance with the AlexNet feature on top-5 scheme, and obtains the second-highest performance with the AlexNet feature on top-1 scheme.

E. Evaluation of the effect of $s$

We need select $s$ samples to constitute the locality-based training set $Y$. This section evaluates the effect of $s$. The LFW face database is utilized in this experiment. The setting is the same as that of Section A. $s$ is set as $0.1, 0.2, ..., 0.8$ times of the number of entries samples of all classes. Figure 2 shows the experiment results. As we can see, the proposed methods obtains good performance when the number of chosen samples $s$ is moderate. The best performance can be obtained when the proportion of chosen samples belongs to $[0.2, 0.5]$.  

VII. CONCLUSION

In this paper, we have proposed a discriminant-based sparse optimization learning model. Based on this model, locality-based discriminant-based sparse representation (LDSR) has been proposed for vision recognition. LDSR obtains the good classification performance because discriminant can help to increase the difference of samples in different classes and to decrease the difference of samples within the same class. Moreover, kernel locality-based discriminant-based sparse representation (KLDSR) was further proposed based on the kernel techniques. KLDSR can take advantages of kernel techniques, discriminant and sparse. To demonstrate the performance of proposed classifiers, extensive experiments have been carried out on several databases including the million-level database: ImageNet. All experiment results prove the effectiveness of the proposed LDSR and KLDSR.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
[2] M. Simon and E. Rodner, “Neural activation constellations: Unsupervised pattern discovery with convolutional networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1143–1151.
[3] O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.
[4] K. P. Murphy, Machine learning: a probabilistic perspective. MIT press, 2012.
[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “Liblinear: A library for large linear classification,” Journal of machine learning research, vol. 9, no. Aug, pp. 1871–1874, 2008.
[6] S. Cai, L. Zhang, W. Zuo, and X. Feng, “A probabilistic collaborative representation based approach for pattern classification.”
[7] 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, vol. 31, no. 2, pp. 210–227, 2009.
[8] A. M. Martinez and A. C. Kak, “Pca versus lda,” IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 2, pp. 228–233, 2001.
[28] I. Naseem, R. Togneri, and M. Bennamoun, “Linear regression for
machine intelligence,” Machine Intell. Mach. Learn. vol. 36, no. 8, pp. 1936.

[29] M. Yang, L. Zhang, X. Feng, and D. Zhang, “Sparse representation
based fisher discrimination dictionary learning for image classification,”
IEEE Trans. Circuits Syst. Video Technol. vol. 21, no. 9, pp. 1329–1337.

[30] R. A. Fisher, “The use of multiple measurements in taxonomic prob-
lems,” Annals of eugenics, vol. 7, no. 2, pp. 179–188, 1936.

[31] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, “Robust face
recognition via adaptive sparse representation,” IEEE Trans. Cybern. vol. 44, no. 12, pp. 2368–2378, 2014.

[32] B. Zhang, A. Perina, V. Murino, and A. Del Bue, “Sparse representation
classification with manifold constraints transfer,” in CVPR, 2015, pp. 4557–4565.

[33] L. Zhang, M. Yang, and X. Deng, “Sparse representation or collaborative
representation: Which helps face recognition?” in 2011 International
Conference on Computer Vision. IEEE, 2011, pp. 471–478.

[34] Y. Chi and F. Porikli, “Classification and boosting with multiple col-
laborative representations,” IEEE transactions on pattern analysis and
machine intelligence, vol. 36, no. 8, pp. 1519–1531, 2014.

[35] “Connecting the dots in multi-class classification: From nearest
subspace to collaborative representation,” in Computer Vision and Pattern
Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 3602–3609.

[36] Y. Xu, Z. Zhong, J. Yang, J. You, and D. Zhang, “A new discriminative
sparse representation method for robust face recognition via l2 regular-
ization,” IEEE Transactions on Neural Networks and Learning Systems,
vol. PP, no. 99, pp. 1–10, 2016.