# Sparse Representation based Fisher Discrimination Dictionary Learning for Image Classification

Meng Yang[a], Lei Zhang[a], Xiangchu Feng[b], and David Zhang[a]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

[b] Department of Applied Mathematics, Xidian University, Xi'an, China

**Abstract:** The employed dictionary plays an important role in sparse representation or sparse coding based image reconstruction and classification, while learning dictionaries from the training data has led to state-of-the-art results in image classification tasks. However, many dictionary learning models exploit only the discriminative information in either the representation coefficients or the representation residual, which limits their performance. In this paper we present a novel dictionary learning method based on the Fisher discrimination criterion. A structured dictionary, whose atoms have correspondences to the subject class labels, is learned, with which not only the representation residual can be used to distinguish different classes, but also the representation coefficients have small within-class scatter and big between-class scatter. The classification scheme associated with the proposed Fisher discrimination dictionary learning (FDDL) model is consequently presented by exploiting the discriminative information in both the representation residual and the representation coefficients. The proposed FDDL model is extensively evaluated on various image datasets, and it shows superior performance to many state-of-the-art dictionary learning methods in a variety of classification tasks.

**Keywords:** dictionary learning, sparse representation, Fisher criterion, image classification

## 1. Introduction

The sparse representation technology has been successfully used in image restoration (Elad and Aharon 2006; Mairal, et al. 2008; Bryt and Elad 2008; Yang, et al. 2008), morphological component analysis (Bobin, et al. 2007) and compressed sensing (Candes 2006). Inspired by the sparse coding mechanism of human vision system (Olshausen and Field 1996; Olshausen and Field 1997), sparse representation represents a signal/image vector as a sparse linear combination of a dictionary of atoms. Recently sparse representation techniques have also led to promising results in face recognition (Wright et al. 2009; Wagner et al. 2009; Yang and Zhang 2010; Yang et al. 2011), handwritten digit and texture classification (Huang and Aviyente 2006; Mairal et al. 2009; Ramirez et al. 2010; Yang et al. 2010; Rodriguez and Sapiro 2007; Mairal et al. 2012), natural image classification (Rodriguez and Sapiro 2007; Yang et al. 2009), and human action recognition (Qiu et al. 2011; Guha and Ward 2012; Wang et al. 2012; Castrodad and Sapiro 2012), etc. In addition, sparse representation, coupled with the low-rank technique, has been used to extract robust features directly from matrix, e.g., the transformation invariant low-rank textures (TILT) (Zhang et al. 2012). The success of sparsity based classification owes to the fact that a high dimensional signal can be sparsely represented by the representative samples of its class in a low dimensional manifold (Wright et al. 2009), while the recent progress of $l_0$- and $l_1$-norm minimization (Tropp and Wright 2010; Yang et al. 2010) facilitates greatly the use of sparse representation to solve large scale problems.

Denote by $\boldsymbol{y}\in\Re^m$ a query sample. The first phase of sparsity based classification is to represent $\boldsymbol{y}$ over a dictionary $\boldsymbol{D} = [\boldsymbol{d}_1,\ldots,\boldsymbol{d}_p]\in\Re^{m\times p}$, i.e., $\boldsymbol{y} \approx \boldsymbol{D\alpha}$, where the representation vector $\boldsymbol{\alpha}\in\Re^p$ has only a few large entries. The following classification phase is based on the solved vector $\boldsymbol{\alpha}$ and the dictionary $\boldsymbol{D}$. The choice of dictionary $\boldsymbol{D}$ is crucial to the success of sparse representation model (Rubinstein et al. 2010). The history of dictionary design could be traced back to 1960s, ranging from the Fast Fourier Transform (FFT) (Cooley and Tukey 1965), Principal Component Analysis without/with missing data (Turk and Pentland 1991; Okatani and Deguchi 2007), wavelets (Mallat 1999), etc., to modern dictionary learning methods, such as Method of Optimal Directions (MOD) (Engan et al. 1999) and KSVD (Aharon et al. 2006). Taking the analytically designed off-the-shelf bases (e.g., FFT, wavelets) as the dictionary (e.g., Huang and Aviyente 2006) is universal to all types of images, but this might not be effective for specific classification tasks such as face recognition. Instead, learning the desired dictionaries from the training data with sparsity regularization has led to state-of-the-art results in image reconstruction (Elad and Aharon 2006; Bryt and Elad 2008; Aharon et al. 2006; Mairal et al. 2012; Zhou et al. 2012) and image classification (Mairal et al. 2009; Zhang and Li 2010; Ramirez et al. 2010; Yang et al. 2010; Yang et al. 2010; Mairal et al. 2008; Rodriguez and Sapiro 2007; Pham and Venkatesh 2008; Jiang et al. 2013; Mairal et al. 2012; Qiu et al. 2011; Jiang et al. 2012; Guha and Ward 2012; Yang et al. 2011; Wang et al. 2012; Castrodad and Sapiro 2012).

The unsupervised dictionary learning (DL) algorithms such as KSVD (Aharon et al. 2006) have achieved promising results in image restoration, but they are not advantageous for image classification tasks because the dictionary is learnt only to faithfully represent the training samples. With the class labels of training samples available, the supervised DL methods exploit the class discrimination information and thus can result in better classification performance. One may use the training samples themselves as the dictionary without learning. For example, Wright et al. (Wright et al. 2009) directly took the training samples of all classes as the dictionary to represent the query face image, and classified it by evaluating which class leads to the minimal reconstruction error of it. The so-called sparse representation based classification (SRC) scheme has shown interesting face recognition results. Nonetheless, the noise and trivial information in the raw training images can make the classification less effective, and the complexity of sparse representation can be very high when the number of training samples is big. In addition, the discriminative information in the training samples is not sufficiently exploited by such a naive supervised DL method. Fortunately, these problems can be addressed, at least to some extent, by learning properly a non-parametric dictionary from the original training samples.

There are mainly two categories of discriminative DL methods for pattern classification. In the first category, a shared dictionary by all classes is learnt but the representation coefficients are discriminative (Mairal et al. 2009; Zhang and Li 2010; Yang et al. 2010; Rodriguez and Sapiro 2007; Pham and Venkatesh 2008; Jiang et al. 2013; Mairal et al. 2012; Lian et al. 2010; Qiu et al. 2011; Jiang et al. 2012). In the DL models proposed by Rodriguez and Sapiro (2007) and Jiang et al. (2013), the samples of the same class are encouraged to have similar sparse representation coefficients. Apart from the $l_0$- or $l_1$-norm sparsity penalty, nonnegative (Hoyer 2002), group (Bengio et al. 2009; Szabo et al. 2011), and structured (Jenatton et al. 2011) sparsity penalty on the representation coefficients have also been proposed in different applications. It is popular to learn a dictionary while training a classifier over the representation coefficients. Mairal et al. (2009) and Pham and Venkatesh (2008) proposed to learn discriminative dictionaries with linear classifiers simultaneously trained. Inspired by the work of Pham and Venkatesh (2008), Zhang and Li (2010) proposed an algorithm called discriminative KSVD (DKSVD) for face recognition, followed by the so-called Label-Consistent KSVD (Jiang et al. 2013). DL from image local features was studied in (Yang et al. 2010; Lian et al. 2010). Recently, Mairal et al. (2012) proposed a task-driven DL framework which minimizes different risk functions of the representation coefficients for different tasks. Generally speaking, the above methods (Mairal et al. 2009; Zhang and Li 2010; Yang et al. 2010; Pham and Venkatesh 2008; Jiang et al. 2013; Mairal et al. 2012; Lian et al. 2010) aim to learn a shared dictionary together with a classifier on the representation coefficients. However, the shared dictionary loses the correspondence between the dictionary atoms and the class labels, and thus performing classification based on the class-specific representation residual is not allowed.

Another category of DL methods learns a dictionary whose atoms have correspondences to the subject class labels (Ramirez et al. 2010; Yang et al. 2010; Mairal et al. 2008; Sprechmann and Sapiro 2010; Wang et al. 2012; Castrodad and Sapiro 2012; Wu et al. 2010). Mairal et al. (2008) introduced a discriminative reconstruction penalty term in the KSVD model (Aharon et al. 2006), and used the learned dictionary for texture segmentation and scene analysis. Yang et al. (2010) and Sprechmann and Sapiro (2010) learned a dictionary for each class with sparse coefficients, and applied it to face recognition and signal clustering, respectively. Castrodad and Sapiro (2012) learned a set of action-specific dictionaries with non-negative penalty on both dictionary atoms and representation coefficients. Wu et al. (2010) learned active basis models from the training images of each category for object detection and recognition. To encourage the dictionaries associated with different classes to be as independent as possible, Ramirez et al. (2010) introduced an incoherence promoting term to the DL model. Based on (Ramirez et al. 2010), Wang et al. (2012) proposed a class-specific DL method for sparse modeling in action recognition. In the above methods (Ramirez et al. 2010; Yang et al. 2010; Mairal et al. 2008; Sprechmann and Sapiro 2010; Wang et al. 2012; Castrodad and Sapiro 2012), the representation residual associated with each class could be used to do classification, but the representation coefficients are not enforced to be discriminative and are not used in the final classification.

Hybrid DL models have also been proposed to learn a shared dictionary and a set of class-specific dictionaries.

Deng et al. (2012) constructed an intra-class face variation dictionary from a generic training dataset, and used it as a shared dictionary to represent the query face image with various variations. Such a method achieves promising performance in face recognition with a single sample per person. Zhou et al. (2012) learned a hybrid dictionary with a Fisher-like regularizer on the representation coefficients, while Kong et al. (2012) learned a hybrid dictionary by introducing an incoherence penalty term to the class-specific sub-dictionaries. Instead of using a flat category structure, Shen et al. (2013) proposed to learn a dictionary with a hierarchical category structure. Although the shared dictionary could make the learned whole hybrid dictionary more compact, how to balance the shared part and the class-specific part in the hybrid dictionary is not a trivial task.

In this paper we propose a Fisher discrimination dictionary learning (FDDL) framework to learn a structured dictionary, i.e., the dictionary atoms have correspondences to the class labels. By FDDL, not only the representation residual associated with each class can be effectively used for classification, the discrimination of representation coefficients will also be exploited. In FDDL, we enforce the sparse representation coefficients having small within-class scatter but big between-class scatter, and enforce each class-specific sub-dictionary having good reconstruction capability to the training samples from that class but poor reconstruction capability to other classes. Therefore, both the representation residual and the representation coefficients of a query sample will be discriminative, and a corresponding classification scheme is proposed to exploit such information. The extensive experiments on various image classification tasks such as face recognition, handwritten digit recognition, gender classification, object categorization and action recognition showed that FDDL could achieve competitive performance with those state-of-the-art DL methods proposed in different tasks.

The rest of this paper is organized as follows. Section 2 briefly reviews some related work. Section 3 presents the proposed FDDL model. Section 4 describes the optimization procedure of FDDL. Section 5 presents the FDDL based classifier. Section 6 conducts extensive experiments, and Section 7 concludes the paper.

## 2. Related Work

### 2.1. Sparse representation based classification

Wright et al. (2009) proposed the sparse representation based classification (SRC) method for robust face recognition (FR). Suppose that there are $K$ classes of subjects, and let $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_K]$ be the set of training samples, where $\boldsymbol{A}_i$ is the subset of training samples from class $i$. Let $\boldsymbol{y}$ be a query sample. The procedures of SRC are summarized as follows.

i.) Sparsely represent $\boldsymbol{y}$ on $\boldsymbol{A}$ via $l_1$-minimization:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha} \right\|_2^2 + \gamma \left\| \boldsymbol{\alpha} \right\|_1 \right\} \tag{1}$$

where $\gamma$ is a scalar constant.

ii.) Perform classification via:

$$\text{identity}\left(\boldsymbol{y}\right) = \arg\min_i \left\{ e_i \right\} \tag{2}$$

where $e_i = \left\| \boldsymbol{y} - \boldsymbol{A}_i \hat{\boldsymbol{\alpha}}_i \right\|_2$, $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_K]$ and $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with class $i$.

Obviously, SRC utilizes the representation residual $e_i$ associated with each class to do classification. Impressive results have been reported in (Wright et al. 2009).

### 2.2. Class-specific dictionary learning

In class-specific dictionary learning (DL), the atoms in the learned dictionary $\boldsymbol{D}$=[$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ..., $\boldsymbol{D}_K$] have class label correspondences to the subject classes, where $\boldsymbol{D}_i$ is the sub-dictionary corresponding to class $i$. After the representation vector $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_K]$ of $\boldsymbol{y}$ is computed, where $\hat{\boldsymbol{\alpha}}_i$ is the sub-vector associated with class $i$, the class-specific representation residual $\left\| \boldsymbol{y} - \boldsymbol{D}_i \hat{\boldsymbol{\alpha}}_i \right\|_2$ could be used for classification. The sub-dictionary $\boldsymbol{D}_i = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{p_i}] \in \Re^{m \times p_i}$ could be learned class by class (Yang et al. 2010; Sprechmann and Sapiro 2010):

$$\min_{\{\boldsymbol{D}_i, \boldsymbol{Z}_i\}} \left\{ \left\| \boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{Z}_i \right\|_F^2 + \lambda \left\| \boldsymbol{Z}_i \right\|_1 \right\} \quad \text{s.t.} \quad \left\| \boldsymbol{d}_j \right\|_2 = 1, \forall j \tag{3}$$

where $\boldsymbol{Z}_i$ is the representation matrix of $\boldsymbol{A}_i$ on $\boldsymbol{D}_i$. Eq. (3) can be seen as the basic model of class-specific DL since each $\boldsymbol{D}_i$ is trained separately from the samples of a specific class.

The basic model in Eq. (3) does not consider the relationship between the sub-dictionaries of different classes. Unlike training the class-specific sub-dictionaries separately in Eq. (3), Ramirez et al. (2010) used an incoherence promoting term to encourage the sub-dictionaries to be as independent as possible. The so-called DL with structured incoherence (DLSI) (Ramirez et al. 2010) is formulated as

$$\min_{\{\boldsymbol{D}_i,\boldsymbol{Z}_i\}} \sum\nolimits_{i=1}^{K} \left\{ \left\|\boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{Z}_i\right\|_F^2 + \lambda \left\|\boldsymbol{Z}_i\right\|_1 \right\} + \eta \sum\nolimits_{i\neq j} \left\|\boldsymbol{D}_i^T \boldsymbol{D}_j\right\|_F^2 \quad \text{s.t.} \ \left\|\boldsymbol{d}_n\right\|_2 = 1, \forall n \tag{4}$$

where the term $\sum\nolimits_{i\neq j} \left\|\boldsymbol{D}_i^T \boldsymbol{D}_j\right\|_F^2$ aims to promote the incoherence between the sub-dictionaries and make the whole class-specific dictionary more distinctive.

## 3. Fisher Discrimination Dictionary Learning (FDDL)

We propose a novel Fisher discrimination dictionary learning (FDDL) scheme, which learns a structured dictionary $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_K]$, where $\boldsymbol{D}_i$ is the sub-dictionary associated with class $i$. By representing a query sample over the learned structured dictionary, the representation residual associated with each class can be naturally employed to classify it, as in the SRC method (Wright et al. 2009). Different from those class-specific DL methods (Ramirez et al. 2010; Yang et al. 2010; Mairal et al. 2008; Sprechmann and Sapiro 2010; Wang et al. 2012; Castrodad and Sapiro 2012; Wu et al. 2010), in FDDL the representation coefficients will also be made discriminative under the Fisher criterion. This will further enhance the discrimination of the dictionary.

Given the training samples $\boldsymbol{A}$=[$\boldsymbol{A}_1$, $\boldsymbol{A}_2$, …, $\boldsymbol{A}_K$] as defined in Section 2.1. Denote by $\boldsymbol{X}$ the sparse representation matrix of $\boldsymbol{A}$ over $\boldsymbol{D}$, i.e., $\boldsymbol{A}\approx\boldsymbol{DX}$. We can write $\boldsymbol{X}$ as $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K]$, where $\boldsymbol{X}_i$ is the representation matrix of $\boldsymbol{A}_i$ over $\boldsymbol{D}$. Apart from requiring that $\boldsymbol{D}$ should have powerful capability to represent $\boldsymbol{A}$ (i.e. $\boldsymbol{A}\approx\boldsymbol{DX}$), we also require that $\boldsymbol{D}$ should have powerful capability to distinguish the images in $\boldsymbol{A}$. To this end, we propose the following FDDL model:

$$J_{(\boldsymbol{D},\boldsymbol{X})} = \arg\min_{(\boldsymbol{D},\boldsymbol{X})} \left\{ r\left(\boldsymbol{A},\boldsymbol{D},\boldsymbol{X}\right) + \lambda_1 \left\|\boldsymbol{X}\right\|_1 + \lambda f\left(\boldsymbol{X}\right) \right\} \quad \text{s.t.} \ \left\|\boldsymbol{d}_n\right\|_2 = 1, \forall n \tag{5}$$

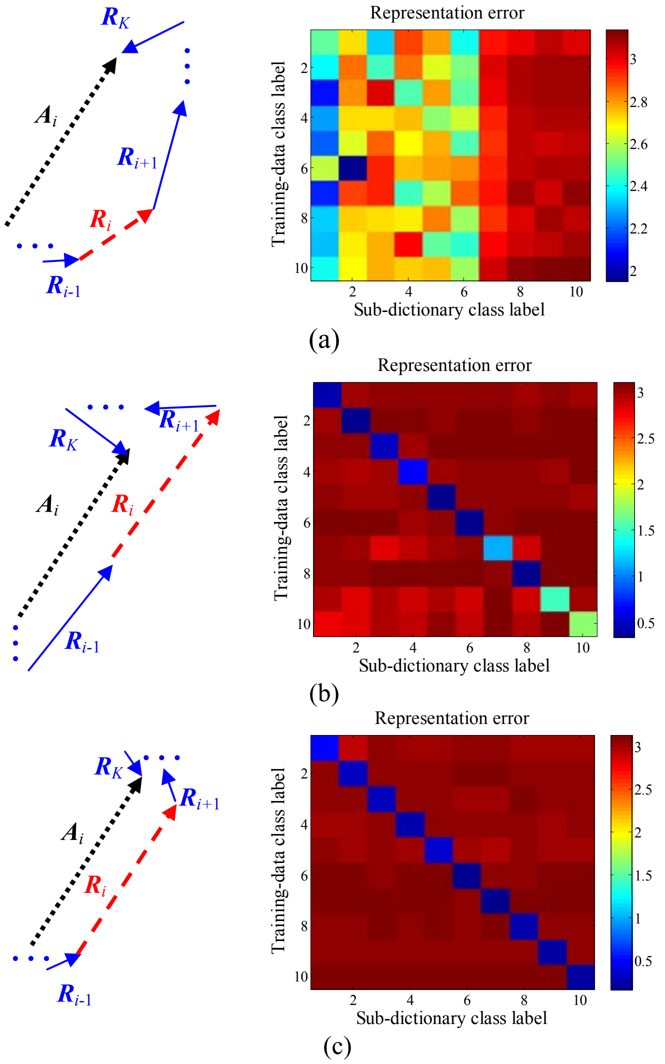
where $r(\boldsymbol{A},\boldsymbol{D},\boldsymbol{X})$ is the discriminative data fidelity term; $\|\boldsymbol{X}\|_1$ is the sparsity penalty; $f(\boldsymbol{X})$ is a discrimination term imposed on the coefficient matrix $\boldsymbol{X}$; and $\lambda_1$ and $\lambda_2$ are scalar parameters. Each atom $\boldsymbol{d}_n$ of $\boldsymbol{D}$ is constrained to have a unit $l_2$-norm to avoid that $\boldsymbol{D}$ has arbitrarily large $l_2$-norm, resulting in trivial solutions of the coefficient matrix $\boldsymbol{X}$. Next let's discuss the design of $r(\boldsymbol{A},\boldsymbol{D},\boldsymbol{X})$ and $f(\boldsymbol{X})$ based on the Fisher discrimination criterion.

### 3.1. Discriminative data fidelity term $r(A,D,X)$

We can write $\boldsymbol{X}_i$ as $\boldsymbol{X}_i=[\boldsymbol{X}_i^1; \ldots; \boldsymbol{X}_i^j; \ldots; \boldsymbol{X}_i^K]$, where $\boldsymbol{X}_i^j$ is the representation coefficients of $\boldsymbol{A}_i$ over $\boldsymbol{D}_j$. Denote by $\boldsymbol{R}_k=\boldsymbol{D}_k\boldsymbol{X}_i^k$ the representation of $\boldsymbol{D}_k$ to $\boldsymbol{A}_i$. First of all, the dictionary $\boldsymbol{D}$ should represent $\boldsymbol{A}_i$ well, and there is $\boldsymbol{A}_i \approx \boldsymbol{D}\boldsymbol{X}_i = \boldsymbol{D}_1\boldsymbol{X}_i^1+\ldots+\boldsymbol{D}_i\boldsymbol{X}_i^i+\ldots+\boldsymbol{D}_K\boldsymbol{X}_i^K=\boldsymbol{R}_1+\ldots+\boldsymbol{R}_i+\ldots+\boldsymbol{R}_K$, where $\boldsymbol{R}_i=\boldsymbol{D}_i\boldsymbol{X}_i^i$. Second, since $\boldsymbol{D}_i$ is associated with the $i^{\text{th}}$ class, it is expected that $\boldsymbol{A}_i$ could be well represented by $\boldsymbol{D}_i$ but not by $\boldsymbol{D}_j$, $j\neq i$. This implies that $\boldsymbol{X}_i^i$ should have some significant coefficients such that $\|\boldsymbol{A}_i\text{-}\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2$ is small, while $\boldsymbol{X}_i^j$ should have very small coefficients such that $\|\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F^2$ is small. Thus we can define the discriminative data fidelity term as

$$r\left(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i\right)=\left\|\boldsymbol{A}_i-\boldsymbol{D}\boldsymbol{X}_i\right\|_F^2+\left\|\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2+\sum\nolimits_{\substack{j=1\\ j\neq i}}^{K}\left\|\boldsymbol{D}_j\boldsymbol{X}_i^j\right\|_F^2 \tag{6}$$

Fig. 1 illustrates the role of the three penalty terms in $r(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i)$. Fig. 1(a) left shows that if we only require $\boldsymbol{D}$ to represent $\boldsymbol{A}_i$ well (i.e., with only the first penalty $\|\boldsymbol{A}_i\text{-}\boldsymbol{D}\boldsymbol{X}_i\|_F^2$), $\boldsymbol{R}_i$ may deviate much from $\boldsymbol{A}_i$ so that $\boldsymbol{D}_i$ could not well represent $\boldsymbol{A}_i$. This problem can be solved by adding the second penalty $\|\boldsymbol{A}_i\text{-}\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2$, as shown in the left of Fig. 1(b). Nonetheless, other sub-dictionaries (for example, $\boldsymbol{D}_{i\text{-}1}$) may also be able to well represent $\boldsymbol{A}_i$, reducing the discrimination capability of $\boldsymbol{D}$. With the third penalty $\|\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F^2$, the representation of $\boldsymbol{D}_j$ to $\boldsymbol{A}_i, j\neq i$, will be small, and the proposed discriminative fidelity term could meet all our expectations, as shown in the left of Fig. 1(c). Let us use a subset of the FRGC 2.0 database to better illustrate the roles of the three terms in Eq. (6). This subset includes 10 subjects with 10 training samples per subject (please refer to Section 6.3 for more information of FRGC 2.0). We learn the dictionary by using the first term, the first two terms and all the three terms, respectively. The representation residuals of the training data over each sub-dictionary are shown in the right column of Fig. 1. One can see that by using only the first term in Eq. (6), we cannot ensure that $\boldsymbol{D}_i$ has the minimal representation residual for $\boldsymbol{A}_i$. By using the first two terms, $\boldsymbol{D}_i$ will have the minimal representation residual for $\boldsymbol{A}_i$ among all sub-dictionaries; however, some training data (e.g., $\boldsymbol{A}_7$, $\boldsymbol{A}_9$, and $\boldsymbol{A}_{10}$) may have big representation residuals over their associated sub-dictionaries because they can be partially represented by other sub-dictionaries. By using all the three terms in Eq. (6), $\boldsymbol{D}_i$ will have not only the minimal but also very small representation residual for $\boldsymbol{A}_i$, while other sub-dictionaries will have big representation residuals of $\boldsymbol{A}_i$.



(a)

(b)

(c)

**Figure 1:** The role of the three penalty terms in $r(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i)$. (a) With only the first term, $\boldsymbol{D}_i$ may not have the minimal representation residual for $\boldsymbol{A}_i$. (b) With the first two terms, $\boldsymbol{D}_i$ will have the minimal representation residual for $\boldsymbol{A}_i$, but some training data (e.g., $\boldsymbol{A}_7$, $\boldsymbol{A}_9$, and $\boldsymbol{A}_{10}$) may have big representation residuals over their associated sub-dictionaries. (c) With all the three terms in Eq. (6), $\boldsymbol{D}_i$ will have not only the minimal but also very small representation residual for $\boldsymbol{A}_i$, while other sub-dictionaries will have big representation residuals of $\boldsymbol{A}_i$.

### 3.2. Discriminative coefficient term $f(\boldsymbol{X})$

To further increase the discrimination capability of dictionary $\boldsymbol{D}$, we can enforce the representation matrix of $\boldsymbol{A}$ over $\boldsymbol{D}$, i.e. $\boldsymbol{X}$, to be discriminative. Based on the Fisher discrimination criterion (Duda et al. 2000), this can be achieved by minimizing the within-class scatter of $\boldsymbol{X}$, denoted by $\boldsymbol{S}_W(\boldsymbol{X})$, and maximizing the between-class scatter of $\boldsymbol{X}$, denoted by $\boldsymbol{S}_B(\boldsymbol{X})$. $\boldsymbol{S}_W(\boldsymbol{X})$ and $\boldsymbol{S}_B(\boldsymbol{X})$ are defined as

$$\boldsymbol{S}_W\left(\boldsymbol{X}\right)=\sum\nolimits_{i=1}^{K}\sum\nolimits_{\boldsymbol{x}_k\in X_i}\left(\boldsymbol{x}_k-\boldsymbol{m}_i\right)\left(\boldsymbol{x}_k-\boldsymbol{m}_i\right)^T \text{ and } \boldsymbol{S}_B\left(\boldsymbol{X}\right)=\sum\nolimits_{i=1}^{K}n_i\left(\boldsymbol{m}_i-\boldsymbol{m}\right)\left(\boldsymbol{m}_i-\boldsymbol{m}\right)^T,$$

where $\boldsymbol{m}_i$ and $\boldsymbol{m}$ are the mean vectors of $\boldsymbol{X}_i$ and $\boldsymbol{X}$, respectively, and $n_i$ is the number of samples in class $\boldsymbol{A}_i$.

The Fisher criterion has been widely used in subspace learning (Wang et al. 2007) to learn a discriminative subspace, and it is usually defined as to minimize the trace ratio $tr(S_W(\boldsymbol{X}))/tr(S_B(\boldsymbol{X}))$, where $tr(\bullet)$ means the trace of a matrix. Instead of minimizing the trace ratio, another commonly used variant of the Fisher criterion is to minimize the trace difference, i.e., minimize $tr(S_W(\boldsymbol{X}))–a\cdot tr(S_B(\boldsymbol{X}))$, where $a$ is a positive constant to balance the contributions of within-class scatter and between-class scatter (Li et al. 2006, Song et al. 2007, Guo et al. 2003, Wang et al. 2007). The relationship between the two types of Fisher criterion has been discussed in detail in (Jia et al. 2009, Wang et al. 2007, Guo et al. 2003). Based on Theorem 1 of (Wang et al. 2007) and Theorem 6 of (Guo *et al.* 2003), the solution of minimizing $tr(S_W(\boldsymbol{X}))–a\cdot tr(S_B(\boldsymbol{X}))$ converges to the solution of minimizing $tr(S_W(\boldsymbol{X}))/tr(S_B(\boldsymbol{X}))$ with a suitable $a$. Since our dictionary learning model contains several other terms apart from the Fisher discrimination term on $\boldsymbol{X}$, we employ the trace difference version of the Fisher criterion, which could make the minimization of the whole FDDL model easier. Meanwhile, we set $a$=1 for simplicity. In Section 6.2 we will show that our model is insensitive to $a$ in a wide range.

Based on the above analysis, we define $f(\boldsymbol{X})$ as $f(\boldsymbol{X})=tr(\boldsymbol{S}_W(\boldsymbol{X}))–tr(\boldsymbol{S}_B(\boldsymbol{X}))$. However, the term $-tr(\boldsymbol{S}_B(\boldsymbol{X}))$ will make $f(\boldsymbol{X})$ non-convex and unstable. To solve this problem, we introduce an elastic term $||\boldsymbol{X}||_F^2$ to $f(\boldsymbol{X})$:

$$f(\boldsymbol{X})= tr(\boldsymbol{S}_W(\boldsymbol{X}))–tr(\boldsymbol{S}_B(\boldsymbol{X}))+\eta||\boldsymbol{X}||_F^2, \tag{7}$$

where $\eta$ is a parameter. The term $||\boldsymbol{X}||_F^2$ could make $f(\boldsymbol{X})$ smoother and convex (the convexity of $f(\boldsymbol{X})$ will be further discussed in Section 4). In addition, in the objective function $J_{(\boldsymbol{D},\boldsymbol{X})}$ (refer to Eq. (5)) of FDDL, there is a sparsity penalty term $||\boldsymbol{X}||_1$. As in elastic-net (Zou and Hastie, 2005), the joint use of $||\boldsymbol{X}||_F^2$ and $||\boldsymbol{X}||_1$ could make the solution of $f(\boldsymbol{X})$ more stable while being sparse.

### 3.3. The whole FDDL model

By incorporating Eqs. (6) and (7) into Eq. (5), we have the following FDDL model:

$$\min\nolimits_{(\boldsymbol{D},\boldsymbol{X})}\left\{\sum\nolimits_{i=1}^{K}r\left(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i\right)+\lambda_1\left\|\boldsymbol{X}\right\|_1+\lambda_2\left(tr\left(\boldsymbol{S}_W\left(\boldsymbol{X}\right)-\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)+\eta\left\|\boldsymbol{X}\right\|_F^2\right)\right\} \text{ s.t. } \left\|\boldsymbol{d}_n\right\|_2=1,\forall n \tag{8}$$

Although the objective function in Eq. (8) is not jointly convex to (***D***, ***X***), we will see that it is convex with respect to each of ***D*** and ***X*** when the other is fixed. Detailed optimization procedures will be presented in Section 4. The dictionary ***D*** to be learned aims to make both the class-specific representation residual and representation coefficients discriminative. Each sub-dictionary $\boldsymbol{D}_i$ will have small representation residuals to the samples from class *i* but have big representation residuals to other classes, while the representation coefficient vectors of samples from one class will be similar to each other but dissimilar to samples from other classes. Such a ***D*** will be very discriminative to classify an input query sample.

A class-specific data representation term was used in (Kong et al. 2012), and a discriminative representation coefficient term was adopted in (Zhou et al. 2012). However, there are much difference between FDDL and these two models. First, both (Kong et al. 2012) and (Zhou et al. 2012) learn a shared dictionary and a set of class-specific sub-dictionaries in their models, while the proposed FDDL only learns a structured dictionary which consists of a set of class-specific sub-dictionaries. Note that although FDDL does not explicitly learn a shared dictionary, it allows across-class representation by using the structured dictionary. Second, FDDL exploits both the representation residual and representation coefficients to learn the discriminative dictionary, while (Kong et al. 2012) and (Zhou et al. 2012) exploit either the representation residual or the representation coefficients in DL.

### 3.4. A simplified FDDL model

The minimization problem in Eq. (8) can be re-formulated as:

$$\begin{aligned}&\min_{\boldsymbol{D},\boldsymbol{X}}\sum_{i=1}^{K}\left(\left\|\boldsymbol{A}_i-\boldsymbol{D}\boldsymbol{X}_i\right\|_F^2+\left\|\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2\right)+\lambda_1\left\|\boldsymbol{X}\right\|_1+\lambda_2\left(tr\left(\boldsymbol{S}_W\left(\boldsymbol{X}\right)-\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)+\eta\left\|\boldsymbol{X}\right\|_F^2\right)\\ &\qquad\qquad\qquad\qquad\text{s.t.}\left\|\boldsymbol{d}_n\right\|_2=1,\forall n;\left\|\boldsymbol{D}_j\boldsymbol{X}_i^j\right\|_F^2\le\varepsilon_f,\forall i\neq j\end{aligned} \tag{9}$$

where $\varepsilon_f$ is a small positive scalar. The constraint $\left\|\boldsymbol{D}_j\boldsymbol{X}_i^j\right\|_F^2\le\varepsilon_f$ guarantees that each class-specific sub-dictionary has poor representation ability for other classes.

It is a little complex to solve the original FDDL model in Eq. (8) or Eq. (9). Considering that $\boldsymbol{X}_i^j$, the representation of $\boldsymbol{A}_i$ over sub-dictionary $\boldsymbol{D}_j$, should be very small for $j\neq i$, we could have a simplified FDDL model by explicitly assuming $\boldsymbol{X}_i^j=0$ for $j\neq i$. In this case, the constraint in Eq. (9) can be well met since $\left\|\boldsymbol{D}_j\boldsymbol{X}_i^j\right\|_F^2=0$

for $j \neq i$. With the simplified FDDL, the representation matrix $\boldsymbol{X}$ becomes block diagonal. The setting of $\boldsymbol{X}_i^j = 0$ will make the within-class scatter $tr(\boldsymbol{S}_W(\boldsymbol{X}))$ small; meanwhile it could be proved that the between-class scatter $tr(\boldsymbol{S}_B(\boldsymbol{X}))$ will be large enough in general (please refer to **Appendix 1** for the proof).

Based on the above discussions, the simplified FDDL model could be written as

$$\min_{\boldsymbol{D},\boldsymbol{X}} \sum\nolimits_{i=1}^{K} \left( \left\| \boldsymbol{A}_i - \boldsymbol{D}\boldsymbol{X}_i \right\|_F^2 + \left\| \boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{X}_i^i \right\|_F^2 \right) + \lambda_1 \left\| \boldsymbol{X} \right\|_1 + \lambda_2 \left( tr\left( \boldsymbol{S}_W\left(\boldsymbol{X}\right) - \boldsymbol{S}_B\left(\boldsymbol{X}\right) \right) + \eta \left\| \boldsymbol{X} \right\|_F^2 \right) \\ \text{s.t.} \left\| \boldsymbol{d}_n \right\|_2 = 1, \forall n; \boldsymbol{X}_i^j = \boldsymbol{0}, \forall i \neq j \tag{10}$$

which could be further formulated as (please refer to **Appendix 2** for the detailed derivation)

$$\min_{(\boldsymbol{D},\boldsymbol{X})} \sum\nolimits_{i=1}^{K} \left( \left\| \boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{X}_i^i \right\|_F^2 + \lambda_1' \left\| \boldsymbol{X}_i^i \right\|_1 + \lambda_2' \left\| \boldsymbol{X}_i^i - \boldsymbol{M}_i^i \right\|_F^2 + \lambda_3' \left\| \boldsymbol{X}_i^i \right\|_F^2 \right) \text{ s.t. } \left\| \boldsymbol{d}_n \right\|_2 = 1, \forall n \tag{11}$$

where $\lambda_1' = \lambda_1/2$, $\lambda_2' = \lambda_2\left(1+\kappa_i\right)/2$, $\kappa_i$=1-$n_i/n$, and $\lambda_3' = \lambda_2\left(\eta - \kappa_i\right)/2$; $\boldsymbol{M}_i^i$ is the mean vector matrix (by taking the mean vector $\boldsymbol{m}_i^i$ as its column vectors) of class $i$, and $\boldsymbol{m}_i^i$ is the column mean vector of $\boldsymbol{X}_i^i$. Clearly, the learning of dictionaries in the simplified FDDL model could be performed class by class.

Compared with the original FDDL model in Eq. (8), the simplified FDDL model in Eq. (11) does not explicitly consider the discrimination between different classes. There are two common ways to improve the discrimination of a classification model: reduce the within-class variation, and enlarge the between-class distance. The FDDL model considers both, while the simplified FDDL model only reduces the within-class variation to enhance the discrimination capability. Fortunately, a large between-class scatter can be guaranteed by simplified FDDL in general, as we proved in **Appendix 1**.

## 4. Optimization of FDDL

We first present the minimization procedure of the original FDDL model in Eq. (8), and then present the solution of the simplified FDDL model in Eq. (11). The objective function in Eq. (8) can be divided into two sub-problems by optimizing $\boldsymbol{D}$ and $\boldsymbol{X}$ alternatively: updating $\boldsymbol{X}$ with $\boldsymbol{D}$ fixed, and updating $\boldsymbol{D}$ with $\boldsymbol{X}$ fixed. The alternative optimization is iteratively implemented to find the desired dictionary $\boldsymbol{D}$ and coefficient matrix $\boldsymbol{X}$.

### 4.1. Update of $X$

Suppose that the dictionary $\boldsymbol{D}$ is fixed, and then the objective function in Eq. (8) is reduced to a sparse representation problem to compute $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K]$. We can compute $\boldsymbol{X}_i$ class by class. When compute $\boldsymbol{X}_i$, all $\boldsymbol{X}_j, j\neq i$, are fixed. The objective function in Eq. (8) is further reduced to:

$$\min_{\boldsymbol{X}_i}\left\{r(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i)+\lambda_1\left\|\boldsymbol{X}_i\right\|_1+\lambda_2 f_i(\boldsymbol{X}_i)\right\} \tag{12}$$

with

$$f_i\left(\boldsymbol{X}_i\right)=\left\|\boldsymbol{X}_i-\boldsymbol{M}_i\right\|_F^2-\sum\nolimits_{k=1}^{K}\left\|\boldsymbol{M}_k-\boldsymbol{M}\right\|_F^2+\eta\left\|\boldsymbol{X}_i\right\|_F^2,$$

where $\boldsymbol{M}_k$ and $\boldsymbol{M}$ are the mean vector matrices (by taking the mean vector $\boldsymbol{m}_k$ or $\boldsymbol{m}$ as all the column vectors) of class $k$ and all classes, respectively. It can be proved that if $\eta>\kappa_i$, $f_i(\boldsymbol{X}_i)$ is strictly convex to $\boldsymbol{X}_i$ (please refer to **Appendix 3** for the proof), where $\kappa_i=1-n_i/n$, $n_i$ and $n$ are the numbers of training samples in the $i^{\text{th}}$ class and all classes, respectively. In this paper, we set $\eta=1$ for simplicity. One can see that all the terms in Eq. (12), except for $\|\boldsymbol{X}\|_1$, are differentiable. We rewrite Eq. (12) as

$$\min_{\boldsymbol{X}_i}\left\{Q\left(\boldsymbol{X}_i\right)+2\tau\left\|\boldsymbol{X}_i\right\|_1\right\} \tag{13}$$

where $Q\left(\boldsymbol{X}_i\right)=r\left(\boldsymbol{A}_i,\boldsymbol{D},\boldsymbol{X}_i\right)+\lambda_2 f_i\left(\boldsymbol{X}_i\right)$, and $\tau=\lambda_1/2$. Let $\tilde{\boldsymbol{X}}_i=\left[\boldsymbol{x}_{i,1}^T,\boldsymbol{x}_{i,2}^T,\cdots,\boldsymbol{x}_{i,n_i}^T\right]^T$, where $\boldsymbol{x}_{i,k}$ is the $k^{\text{th}}$ column vector of matrix $\boldsymbol{X}_i$. Because $Q(\boldsymbol{X}_i)$ is strictly convex and differentiable to $\boldsymbol{X}_i$, the Iterative Projection Method (IPM) (Rosasco et al. 2009, whose speed could be improved by FISTA (Beck and Teboulle 2009)) can be employed to solve Eq. (13), as described in Table 1.

The update of representation matrix $\boldsymbol{X}$ in the simplified FDDL model (i.e., Eq. (11)) is a special case of that in FDDL with $Q\left(\boldsymbol{X}_i\right)=\left\|\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2+\lambda_2'\left\|\boldsymbol{X}_i^i-\boldsymbol{M}_i^i\right\|_2^2+\lambda_3'\left\|\boldsymbol{X}_i^i\right\|_F^2$ and $\boldsymbol{X}_i^j=0$ for $j\neq i$, which could also be efficiently solved by the algorithm in Table 1. In simplified FDDL, we set $\eta=\kappa_i=1-n_i/n$ (i.e., $\lambda_3'=0$) and in this case $Q(\boldsymbol{X}_i)$ is convex w.r.t. $\boldsymbol{X}_i$.

**Table 1:** The update of representation matrix ***X*** in FDDL.

```
Algorithm of updating $\boldsymbol{X}$ in FDDL
1. Input: $\sigma$, $\tau$ >0.
2. Initialization: $\tilde{\boldsymbol{X}}_i^{(1)} = \mathbf{0}$ and $h$=1.
3. While convergence or the maximal iteration number is not reached do
       $h = h+1$
       $\tilde{\boldsymbol{X}}_i^{(h)} = \boldsymbol{S}_{\tau/\sigma}\left( \tilde{\boldsymbol{X}}_i^{(h-1)} - \frac{1}{2\sigma} \nabla Q\left( \tilde{\boldsymbol{X}}_i^{(h-1)} \right) \right)$        (14)
       where $\nabla Q\left( \tilde{\boldsymbol{X}}_i^{(h-1)} \right)$ is the derivative of $Q(\boldsymbol{X}_i)$ w.r.t. $\tilde{\boldsymbol{X}}_i^{(h-1)}$, and $\boldsymbol{S}_{\tau/\sigma}$ is a
       component-wise soft thresholding operator defined by (Wright et al. 2009a):
       $\left[ \boldsymbol{S}_{\tau/\sigma}(\boldsymbol{x}) \right]_j = \begin{cases} 0 & \left| x_j \right| \le \tau/\sigma \\ x_j - \text{sign}\left( x_j \right) \tau/\sigma & \text{otherwise} \end{cases}.$
4. Return $\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{X}}_i^{(h)}$.
```

### 4.2. Update of *D*

Let's then discuss how to update ***D*** = [***D***$_1$, ***D***$_2$, …, ***D***$_K$] when ***X*** is fixed. We also update $\boldsymbol{D}_i = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_{p_i}]$ class by class. When update ***D***$_i$, all ***D***$_j$, $j \neq i$, are fixed. The objective function in Eq. (8) is reduced to:

$$\min_{\boldsymbol{D}_i} \left\{ \left\| \hat{\boldsymbol{A}} - \boldsymbol{D}_i \boldsymbol{X}^i \right\|_F^2 + \left\| \boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{X}_i^i \right\|_F^2 + \sum\nolimits_{j=1, j \neq i}^{K} \left\| \boldsymbol{D}_i \boldsymbol{X}_j^i \right\|_F^2 \right\} \text{ s.t. } \left\| \boldsymbol{d}_l \right\|_2 = 1, l = 1, \cdots, p_i \tag{15}$$

where $\hat{\boldsymbol{A}} = \boldsymbol{A} - \sum\nolimits_{j=1, j \neq i}^{K} \boldsymbol{D}_j \boldsymbol{X}^j$ and $\boldsymbol{X}^i$ is the representation matrix of ***A*** over ***D***$_i$. Eq. (15) could be re-written as

$$\min_{\boldsymbol{D}_i} \left\| \boldsymbol{\Lambda}_i - \boldsymbol{D}_i \boldsymbol{Z}_i \right\|_F^2 \text{ s.t. } \left\| \boldsymbol{d}_l \right\|_2 = 1, l = 1, \cdots, p_i \tag{16}$$

where $\boldsymbol{\Lambda}_i = \left[ \hat{\boldsymbol{A}} \; \boldsymbol{A}_i \; \mathbf{0} \cdots \mathbf{0} \; \mathbf{0} \cdots \mathbf{0} \right]$, $\boldsymbol{Z}_i = \left[ \boldsymbol{X}^i \; \boldsymbol{X}_i^i \; \boldsymbol{X}_1^i \cdots \boldsymbol{X}_{i-1}^i \; \boldsymbol{X}_{i+1}^i \cdots \boldsymbol{X}_K^i \right]$, and **0** is a zero matrix with appropriate size based on the context. Eq. (16) can be efficiently solved by updating each dictionary atom one by one via the algorithm like (Yang et al. 2010) or (Mairal et al. 2008).

The update of dictionary in simplified FDDL is the same as original FDDL except that Eq. (16) becomes a simpler one with $\boldsymbol{\Lambda}_i = \boldsymbol{A}_i$ and $\boldsymbol{Z}_i = \boldsymbol{X}_i^i$.

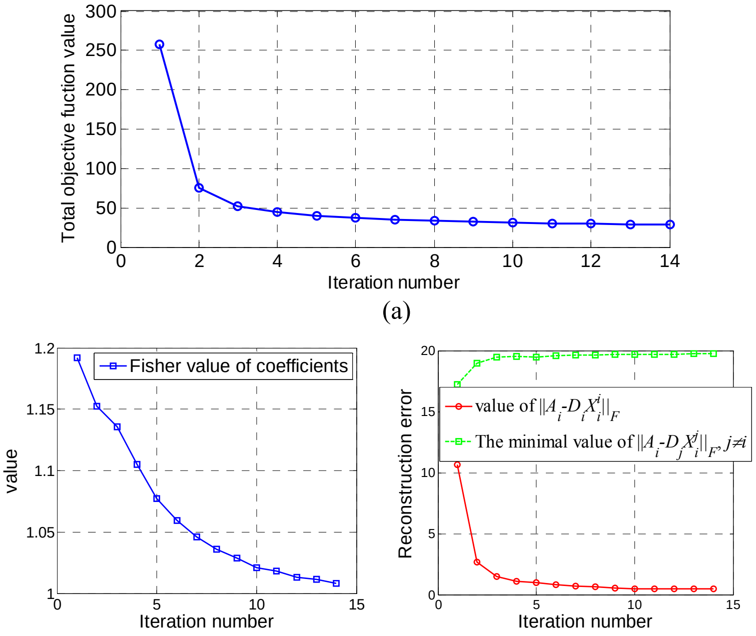
### 4.3. Algorithm of FDDL

The complete algorithm of FDDL is summarized in Table 2. The algorithm converges since the cost function in Eq. (8) or Eq. (11) is lower bounded and can only decrease in the two alternative minimization stages (i.e., updating $\boldsymbol{X}$ and updating $\boldsymbol{D}$). An example of FDDL minimization is shown in Fig. 2 by using the Extended Yale B face database (Georghiades et al. 2001). Fig. 2(a) illustrates the convergence of FDDL. Fig. 2(b) shows that the Fisher ratio $tr(\boldsymbol{S}_W(\boldsymbol{X}))/tr(\boldsymbol{S}_B(\boldsymbol{X}))$, which is basically equivalent to $tr(\boldsymbol{S}_W(\boldsymbol{X}))-tr(\boldsymbol{S}_B(\boldsymbol{X}))$ in characterizing the discrimination capability of $\boldsymbol{X}$, decreases with the increase of iteration number. This indicates that the coefficients $\boldsymbol{X}$ are discriminative by the proposed FDDL algorithm. Fig. 2(c) plots the curves of $\|\boldsymbol{A}_i\text{-}\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F$ ($i$=10 here) and the minimal value of $\|\boldsymbol{A}_i\text{-}\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F$, $j$=1,2,…,$K$, $j\neq i$, showing that $\boldsymbol{D}_i$ represents $\boldsymbol{A}_i$ well, but $\boldsymbol{D}_j$, $j\neq i$, has poor representation ability to the samples in $\boldsymbol{A}_i$.

**Table 2:** Algorithm of Fisher discrimination dictionary learning.

**Fisher Discrimination Dictionary Learning (FDDL)**

1. **Initialize *D*.**
   We initialize the atoms of $\boldsymbol{D}_i$ as the eigenvectors of $\boldsymbol{A}_i$.
2. **Update coefficients *X*.**
   Fix $\boldsymbol{D}$ and solve $\boldsymbol{X}_i$, $i$=1,2,…,$K$, one by one by solving Eq. (13) with the algorithm in **Table 1**.
3. **Update dictionary *D*.**
   Fix $\boldsymbol{X}$ and update each $\boldsymbol{D}_i$, $i$=1,2,…,$K$, by solving Eq. (16) :
   1) Let $\boldsymbol{Z}_i=\left[z_1;z_2;\cdots;z_{p_i}\right]$ and $\boldsymbol{D}_i=\left[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_{p_i}\right]$, where $z_j$, $j$=1,2,…,$p_i$, is the row vector of $\boldsymbol{Z}_i$, and $\boldsymbol{d}_j$ is the $j^{\text{th}}$ column vector of $\boldsymbol{D}_i$.
   2) Fix all $\boldsymbol{d}_l, l\neq j$, update $\boldsymbol{d}_j$. Let $\boldsymbol{Y}=\boldsymbol{A}_i-\sum_{l\neq j}\boldsymbol{d}_l z_l$ . The minimization of Eq. (16) becomes
   $$\min_{\boldsymbol{d}_j}\left\|\boldsymbol{Y}-\boldsymbol{d}_j\boldsymbol{z}_j\right\|_F^2 \text{ s.t. } \left\|\boldsymbol{d}_j\right\|_2=1;$$
   After some deviation (Yang et al. 2010), we could get the solution $\boldsymbol{d}_j=\boldsymbol{Y}\boldsymbol{z}_j^T\Big/\left\|\boldsymbol{Y}\boldsymbol{z}_j^T\right\|_2$ .
   3) Using the above procedures, we can update all $\boldsymbol{d}_j$, and hence the whole dictionary $\boldsymbol{D}_i$ is updated.
4. **Output.**
   Return to **step 2** until the objective function values in adjacent iterations are close enough or the maximum number of iterations is reached. Then output $\boldsymbol{X}$ and $\boldsymbol{D}$.



(a)

(b)

(c)

**Figure 2**: An example of FDDL minimization process on the Extended Yale B face database. (a) The convergence of FDDL. (b) The curve of Fisher ratio $tr(\boldsymbol{S}_W(\boldsymbol{X}))/tr(\boldsymbol{S}_B(\boldsymbol{X}))$ versus the iteration number. (c) The curves of the reconstruction residual of $\boldsymbol{D}_i$ to $\boldsymbol{A}_i$ and the minimal reconstruction residual of $\boldsymbol{D}_j$ to $\boldsymbol{A}_i$, $j\neq i$, versus the iteration number.

### 4.4. Time complexity

In the proposed FDDL algorithm, the update of coding coefficients for each sample is a sparse coding problem, whose time complexity is approximately $O(q^2p^\varepsilon)$ (Kim et al. 2007, Nesterov and Nemirovskii 1994), where $\varepsilon\geq1.2$ is a constant, $q$ is the feature dimensionality and $p$ is the number of dictionary atoms. So the total time complexity of updating coding coefficients in FDDL is $nO(q^2p^\varepsilon)$, where $n$ is the total number of training samples. The time complexity of updating dictionary atoms (i.e., Eq. (16)) is $\Sigma_i p_i O(2nq)$, where $p_i$ is the number of dictionary atoms in $\boldsymbol{D}_i$. Therefore, the overall time complexity of FDDL is approximately $\vartheta(nO(q^2p^\varepsilon)+ \Sigma_i p_i O(2nq))$, where $\vartheta$ is the total number of iterations.

For simplified FDDL, the time complexity of updating coding coefficients is $\Sigma_i n_i O(q^2p_i^\varepsilon)$, where $n_i$ is the number training samples in the $i^{\text{th}}$ class. The time complexity of updating dictionary atoms is $\Sigma_i p_i O(n_i q)$. Therefore, the overall time complexity of simplified FDDL is $\vartheta(\Sigma_i n_i O(q^2p_i^\varepsilon)+ \Sigma_i p_i O(n_i q))$. Since $n$=$\Sigma_i n_i$ and $p$=$\Sigma_i p_i$, we can see that the simplified FDDL algorithm has much lower time complexity than the original FDDL algorithm.

Let's evaluate the running time of FDDL and simplified FDDL by using a subset of FRGC 2.0 with 316 subjects (5 training samples per subject, please refer to Section 6.3 for more detailed experimental setting). We also report the running time of shared dictionary learning method DKSVD (Zhang and Li 2010), class-specific dictionary learning method DLSI (Ramirez et al. 2010), and hybrid dictionary learning method COPAR (Kong and Wang 2012). The iteration number of all dictionary learning methods is set as 20. Under the MATLAB R2011a programming environment and in a desktop of 2.90GHZ with 4.00GB RAM, the running time of FDDL and simplified FDDL is 627.6s and 31.2s, respectively, while the running time of DKSVD, DLSI and COPAR is 728.5s, 1000.6s and 5708.8s, respectively.

## 5. The Classification Scheme

Once the dictionary $\boldsymbol{D}$ is learned, it could be used to represent a query sample $\boldsymbol{y}$ and judge its label. According to how the dictionary $\boldsymbol{D}$ is learned, different information can be utilized to perform the classification task. In (Mairal et al. 2009; Zhang and Li 2010; Yang et al. 2010; Pham and Venkatesh 2008; Jiang et al. 2013; Mairal et al. 2012; Lian et al. 2010; Jiang et al. 2012), a shared dictionary by all classes is learned, and the sparse representation coefficients are used for classification. In the SRC scheme (Wright et al. 2009), the original training samples are employed as a structured dictionary to represent the query sample, and the representation residual by each class is used for classification. In (Ramirez et al. 2010; Mairal et al. 2008; Wang et al. 2012; Castrodad and Sapiro 2012), the query sample is sparsely coded on each class-specific sub-dictionary, and the representation residual is computed for classification. With the proposed FDDL scheme, however, both the representation residual and the representation coefficients will be discriminative, and hence we can make use of both of them to achieve more accurate classification results.

By FDDL, not only the desired dictionary $\boldsymbol{D}$ is learned from the training dataset $\boldsymbol{A}$, the representation matrix $\boldsymbol{X}_i$ of each class $\boldsymbol{A}_i$ is also computed. With $\boldsymbol{X}_i$, the mean coefficient vector of class $\boldsymbol{A}_i$, denoted by $\boldsymbol{m}_i$, could be calculated. (For simplified FDDL, the mean coefficient vector for each class can be constructed by $\boldsymbol{m}_i=\left[\mathbf{0};\cdots;\boldsymbol{m}_i^i;\cdots;\mathbf{0}\right]$, where $\boldsymbol{m}_i^i$ is the mean vector of $\boldsymbol{X}_i^i$.) The mean vector $\boldsymbol{m}_i$ can be viewed as the center of class $\boldsymbol{A}_i$ in the transformed space spanned by the dictionary $\boldsymbol{D}$. In FDDL, not only the class-specific sub-dictionary

$\boldsymbol{D}_i$ is forced to represent the training samples in $\boldsymbol{A}_i$, the representation coefficient vectors in $\boldsymbol{X}_i$ are also forced to be close to $\boldsymbol{m}_i$ and be far from $\boldsymbol{m}_j$, $j \neq i$. Suppose that the query sample $\boldsymbol{y}$ is from class $\boldsymbol{A}_i$, then its representation residual by $\boldsymbol{D}_i$ will be small, while its representation vector over $\boldsymbol{D}$ will be more likely close to $\boldsymbol{m}_i$. Therefore, the mean vectors $\boldsymbol{m}_i$ can be naturally employed to improve the classification performance. According to the number of training samples per class, here we propose two classifiers, the global classifier (GC) and local classifier (LC), which are described as follows.

### 5.1. Global classifier

When the number of training samples per class is relatively small, the learned sub-dictionary $\boldsymbol{D}_i$ may not be able to faithfully represent the query samples of this class, and hence we represent the query sample $\boldsymbol{y}$ over the whole dictionary $\boldsymbol{D}$. On the other hand, in the test stage the $l_1$-norm regularization on the representation coefficient may be relaxed to $l_2$-norm regularization for faster speed, as discussed in (Zhang et al. 2011). With these considerations, we use the following global representation model:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \left\{ \left\| \boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha} \right\|_2^2 + \gamma \left\| \boldsymbol{\alpha} \right\|_p \right\} \tag{17}$$

where $\gamma$ is a constant and $\|\bullet\|_p$ means $l_p$-norm, $p$=1 or 2. Note that when $p$=2, an analytical regularized least square solution to $\hat{\boldsymbol{\alpha}}$ can be readily obtained so that the representation process is extremely fast (Zhang et al. 2011).

Denote by $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_K]$, where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient sub-vector associated with sub-dictionary $\boldsymbol{D}_i$. In the training stage of FDDL, we have enforced the class-specific representation residual to be discriminative. Therefore, if $\boldsymbol{y}$ is from class $i$, the residual $\left\| \boldsymbol{y} - \boldsymbol{D}_i\hat{\boldsymbol{\alpha}}_i \right\|_2^2$ should be small while $\left\| \boldsymbol{y} - \boldsymbol{D}_j\hat{\boldsymbol{\alpha}}_j \right\|_2^2$, $j \neq i$, should be big. In addition, the representation vector $\hat{\boldsymbol{\alpha}}$ should be close to $\boldsymbol{m}_i$ but far from the mean vectors of other classes. By considering the discrimination capability of both representation residual and representation vector, we could define the following metric for classification:

$$e_i = \left\| \boldsymbol{y} - \boldsymbol{D}_i\hat{\boldsymbol{\alpha}}_i \right\|_2^2 + w \cdot \left\| \hat{\boldsymbol{\alpha}} - \boldsymbol{m}_i \right\|_2^2 \tag{18}$$

where $w$ is a preset weight to balance the contribution of the two terms to classification. The classification rule is simply set as $\text{identity}(\boldsymbol{y}) = \arg\min_i \{e_i\}$.

### 5.2. Local classifier

When the number of training samples of each class is relatively large, the sub-dictionary $\boldsymbol{D}_i$ is able to well span the subspace of class *i*. In this case, to reduce the interference from other sub-dictionaries and to reduce the complexity of sparse representation, we can represent $\boldsymbol{y}$ locally over each sub-dictionary $\boldsymbol{D}_i$ instead of the whole dictionary $\boldsymbol{D}$; that is, $\hat{\boldsymbol{\alpha}}_i$=argmin$_{\alpha_i}$ $\|\boldsymbol{y}-\boldsymbol{D}_i\boldsymbol{\alpha}_i\|_2^2+\gamma\|\boldsymbol{\alpha}_i\|_p$. However, since in the dictionary learning stage we have forced the representation vectors of $\boldsymbol{A}_i$ over $\boldsymbol{D}_i$ to be close to their mean, i.e., $\boldsymbol{m}_i^i$, in the test stage we can also force the representation vector of query sample $\boldsymbol{y}$ over $\boldsymbol{D}_i$ to be close to $\boldsymbol{m}_i^i$ so that the representation process can be more informative. With the above considerations, we propose the following local representation model:

$$\hat{\boldsymbol{\alpha}}_i=\arg\min_{\alpha_i}\left\{\|\boldsymbol{y}-\boldsymbol{D}_i\boldsymbol{\alpha}_i\|_2^2+\gamma_1\|\boldsymbol{\alpha}_i\|_p+\gamma_2\left\|\boldsymbol{\alpha}_i-\boldsymbol{m}_i^i\right\|_2^2\right\} \tag{19}$$

where $\gamma_1$ and $\gamma_2$ are constants. Again, when $p$=2, an analytical solution to $\hat{\boldsymbol{\alpha}}_i$ can be obtained. Based on the representation model in Eq. (19), the metric used for classification can be readily defined as:

$$e_i=\|\boldsymbol{y}-\boldsymbol{D}_i\hat{\boldsymbol{\alpha}}_i\|_2^2+\gamma_1\|\hat{\boldsymbol{\alpha}}_i\|_p+\gamma_2\left\|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{m}_i^i\right\|_2^2 \tag{20}$$

and the final classification rule is still $\text{identity}(\boldsymbol{y})=\arg\min_i\{e_i\}$.

## 6. Experimental Results

We verify the performance of FDDL on various image classification tasks. Section 6.1 discusses the model and parameter selection; Section 6.2 illustrates the effectiveness of FDDL in improving the Fisher discrimination criterion of representation coefficients; Sections 6.3~6.7 perform experiments on face recognition, handwritten digit recognition, gender classification, object categorization and action recognition, respectively.

### 6.1. Model and parameter selection

We discuss the various issues involved in the proposed scheme, including DL model selection (i.e., FDDL or simplified FDDL), classification model selection (e.g., GC or LC), the number of dictionary atoms, $l_1$-norm or $l_2$-norm regularization, and parameter selection. In studying the DL model selection, the parameters $\gamma$ and $w$ in GC and parameters $\gamma_1$ and $\gamma_2$ in LC are predefined. Specifically, we select the values of $\gamma$ and $\gamma_1$ from set {0.001, 0.01, 0.1}, and select the values of $w$ and $\gamma_2$ from set {0, 0.001}. Given a dictionary learning and classification model, we report the best performance of different classifiers.

### *6.1.1 Model selection in dictionary learning and classification*

We first discuss the classification model selection. As analyzed in Sections 5.1 and 5.2, the GC and LC are suitable for small-sample-size problem and enough-training-sample problem, respectively. In the experiments of this sub-section, the $l_1$-norm regularization is used when representing a query sample.

The FR rates of FDDL and simplified FDDL coupled with GC and LC on the AR database (Martinez and Benavente 1998) are listed in Table 3 (more information about the experiment settings can be found in Section 6.3). Here we set $\lambda_1$=0.005 and $\lambda_2$=0.01 in the DL stage. It can be seen that GC achieves much better performance than LC with about 20% advancement. This validates that when the number of training samples per class (denoted by $N_{ts}$) is not sufficient and different classes share some similarities, the cross-class representation in GC is helpful to represent the test sample. The competition of different classes in the representation process makes the representation residual discriminative for classification.

**Table 3:** FR rates of FDDL and simplified FDDL coupled with GC or LC on the AR database.

| $N_{ts}$ | 4 | | 7 | |
|---|---|---|---|---|
| | GC | LC | GC | LC |
| FDDL | **86.3%** | 61.5% | 92.6% | 74.8% |
| Simplified FDDL | 86.0% | 61.4% | **93.0%** | 74.8% |

**Table 4**: Performance of FDDL and simplified FDDL coupled with GC or LC in USPS digit recognition.

| $N_{ts}$ | 5 | | 10 | | 100 | | 300 | |
|---|---|---|---|---|---|---|---|---|
| | GC | LC | GC | LC | GC | LC | GC | LC |
| FDDL | 78.9% | **79.8%** | 82.9% | **84.3%** | 90.2% | 94.1% | 90.8% | 94.1% |
| Simplified FDDL | 78.5% | 79.5% | 82.9% | 84.1% | 92.9% | **94.2%** | 94.3% | **95.0%** |

**Table 5**: Performance of SRC, simplified FDDL and FDDL in FR on Multi-PIE.

| $N_{ts}$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| SRC | 64.8% | 75.4% | 79.0% | 90.4% | 95.2% | 97.0% |
| Simplified FDDL (LC) | 51.3% | 53.4% | 55.2% | 76.2% | **95.2%** | **98.7%** |
| FDDL (LC) | **61.9%** | **72.8%** | **76.5%** | **88.6%** | 95.1% | 98.4% |
| Simplified FDDL (GC) | 70.7% | 84.8% | 88.7% | 97.3% | **98.3%** | 99.3% |
| FDDL (GC) | **71.9%** | **86.6%** | **91.1%** | **97.6%** | 98.2% | **99.4%** |

Table 4 lists the recognition rates of FDDL and simplified FDDL on the USPS handwritten digit dataset (Hull, J.J. 1994) (more information about the experiment settings can be found in Section 6.4). We set $\lambda_1$=0.05 and $\lambda_2$=0.005. In this experiment, when the number of training samples is not big (e.g., 5 or 10), all methods have similar performance; when the number of training samples is relatively large (e.g., 100 and 300), LC outperforms GC. This shows that LC is more powerful than GC when there are sufficient training samples per class.

We then compare the performance of simplified FDDL and FDDL. As mentioned in Section 3.4, simplified FDDL is a special but very useful case of FDDL by assuming $\boldsymbol{X}_i^j = \boldsymbol{0}$, $j\neq i$, when we represent the training data $\boldsymbol{A}$ = [$\boldsymbol{A}_1$, $\boldsymbol{A}_2$, …, $\boldsymbol{A}_K$] over the structured dictionary $\boldsymbol{D}$=[$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ..., $\boldsymbol{D}_K$]. In many situations, different classes have a similar number of training samples and have similar variations (i.e., $\boldsymbol{A}_i$ are "balanced"), and this assumption holds well because $\boldsymbol{A}_i$ tends to be represented mostly by $\boldsymbol{D}_i$. Actually, in Tables 3 and 4, we have seen that simplified FDDL achieves similar recognition rates to FDDL. However, this assumption may not be satisfied when $\boldsymbol{A}_i$ are less balanced, and in such cases FDDL can perform much better than simplified FDDL.

We conduct a face recognition experiment on the Multi-PIE database (please refer to Section 6.3 for more information of Multi-PIE). The face images of 60 subjects with illumination and expression variations (including neutral, smile in session 1 and session 3, surprise and squint in section 2) are used. Each subject has $N_{ts}$ training samples and 27 testing samples (15 samples with illumination variations and 12 samples with expression variations). When $N_{ts} \leq 5$, the training samples of the first 30 subjects have expression variations, while the training samples of the last 30 subjects have illumination variations. When $N_{ts} > 5$, we add face images with illumination variations to the first 30 subjects and add face images with expression variations to the last 30 subjects. We set $\lambda_1$=$\lambda_2$=0.01 in FDDL and simplified FDDL.

Table 5 lists the FR results of simplified FDDL and FDDL. We also report the results of SRC for reference. From Table 5 we can see that FDDL is visibly better than its simplified version when $N_{ts} \leq 5$. If LC is used, the advantage of FDDL over simplified FDDL is more obvious. This is because in this experiment when $N_{ts} \leq 5$ the training samples in one class are limited in characterizing the different face variations, and thus the assumption $\boldsymbol{X}_i^j = \boldsymbol{0}$, $j\neq i$, does not hold very well. Fortunately, FDDL could exploit the complementary information from different classes to learn the dictionary. In contrast, simplified FDDL ignores the collaborative representation between different classes, degrading its performance. When $N_{ts} > 5$, different classes become more balanced in illumination and expression variations, and the assumption $\boldsymbol{X}_i^j = \mathbf{0}$, $j \neq i$, can be well satisfied. As a result, the performance of simplified FDDL becomes close to FDDL. In addition, both simplified FDDL and FDDL perform much better than SRC under different $N_{ts}$.

In our following experiments, including face recognition, digit recognition, gender classification, object categorization and action recognition, all classes have similar number of training samples and similar variations. Based on our above analyses, we adopt simplified FDDL to learn dictionaries except for face recognition (the training samples in face recognition are usually less insufficient and FDDL performs slightly better than simplified FDDL). In the classification phase, GC is used in face recognition, object categorization and action recognition, LC is used in handwritten digit recognition, while both GC and LC are tested in gender classification.

*6.1.2 The number of dictionary atoms*

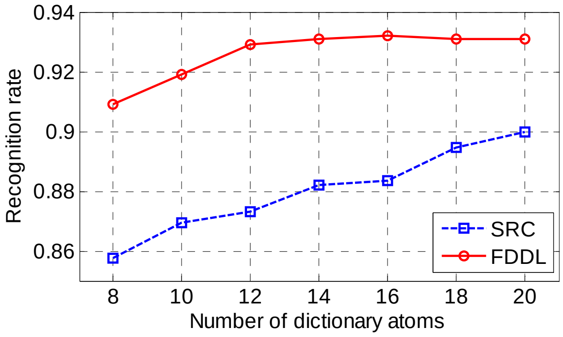


**Figure 3**: The recognition rates of FDDL and SRC versus the number of dictionary atoms.

One important parameter in FDDL is the number of atoms in $\boldsymbol{D}_i$, denoted by $p_i$. Usually, we let all $p_i$ equal, $i$=1,2,…,$K$. Here we use SRC as the baseline method to analyze the effect of $p_i$ on the performance of FDDL by using face recognition experiment on the Extended Yale B database, which consists of 2,414 frontal-face images from 38 individuals. For each subject, we randomly select 20 images for training, with the others (about 44 images per subject) for testing. Because SRC uses the original training samples as dictionary, we randomly select $p_i$ training samples as the dictionary atoms and run 10 times the experiment to compute the average recognition rate. Fig. 3 plots the recognition rates of FDDL and SRC versus the number of dictionary atoms. We can see that in all cases FDDL has at least 3% improvement over SRC. Especially, even with $p_i$=8, FDDL can still have higher recognition rate than SRC with $p_i$=20. Besides, from $p_i$=20 to $p_i$=8, the recognition rate of FDDL drops by 2.2%, compared to 4.2% for SRC. This demonstrates that FDDL is effective to learn a compact and representative dictionary, which can reduce the computational cost and improve the recognition rate simultaneously.

### *6.1.3 $l_1$-norm and $l_2$-norm regularization*

It is indicated in (Zhang et al. 2011) that the $l_1$-norm sparsity penalty on the representation coefficients may not be critical in the SRC classifier. In the DL stage of FDDL, both $l_1$-norm and $l_2$-norm regularizations are imposed on the representation coefficients $\boldsymbol{X}$. In this sub-section, we first evaluate the role of $l_1$-norm regularization (i.e., $||\boldsymbol{X}||_1$) and $l_2$-norm regularization (i.e., $\|\boldsymbol{X}\|_F^2$ ) in Eq. (8) by varying the values of parameters $\lambda_1$ and $\eta$ ($\eta$>$\kappa_i$, where $\kappa_i$=1-$n_i$/$n$), and then evaluate the performance of GC and LC with $l_1$-norm and $l_2$-norm regularizations (i.e., letting $p$=1 and $p$=2 in Eq. (17) and Eq. (19)).

Table 6 lists the recognition rates of FDDL with different values of $\lambda_1$ and $\eta$ on the Extended Yale B database (the experiment setting is the same as that in Section 6.1.2). The GC is used in this experiment and we set $\lambda_2$=0.005. It can be seen that without $l_1$-norm regularization (i.e., $\lambda_1$ =0) in the learning stage of FDDL, the performance will degrade (i.e., FR rate with $\lambda_1$ =0 is lower than that with $\lambda_1$ =0.005). With $l_1$-norm regularization (e.g., $\lambda_1$ =0.005) in the learning stage, varying the strength of $l_2$-norm regularization has little effect on the final classification rate. This finding shows that $l_1$-norm sparse regularization is useful in learning discriminative dictionary for pattern classification.

**Table 6:** FR rates on the Extended Yale B database with various parameter settings of ($\lambda_1$, $\eta$).

| Parameters | (0.005, $\kappa_i$) | (0.005, 1) | (0.005, 5) | (0, 1) | (0, 5) |
|---|---|---|---|---|---|
| $l_1$-regularized GC | **92.4%** | 92.1% | **92.4%** | 90.8% | 91.7% |
| $l_2$-regularized GC | 91.1% | 90.7% | 91.3% | 88.6% | 91.6% |

**Table 7:** Recognition rates on the USPS database with various parameter settings of ($\lambda_1$, $\eta$).

| Parameters | (0.05, $\kappa_i$) | (0.05, 1) | (0.05, 5) | (0, 1) | (0, 5) |
|---|---|---|---|---|---|
| $l_1$-regularized LC | 95.0% | 95.0% | **95.2%** | 93.1% | 93.3% |
| $l_2$-regularized LC | 93.3% | 93.3% | 93.4% | 91.6% | 93.3% |

In Table 6 we can also see that when $\lambda_1$ =0.005, the $l_2$-norm regularized GC has lower recognition rates than the $l_1$-norm regularized GC. This is because $l_1$-norm regularization is used in the DL stage, so that if $l_1$-norm regularization is not employed to represent the query sample, the discrimination ability of the learnt dictionary may not be fully exploited.

We then apply FDDL to the USPS handwritten digit database with 300 training samples per class. The recognition rates are listed in Table 7. The LC classifier is used in this experiment. We fix $\lambda_2$=0.005, and vary the values of $\lambda_1$ and $\eta$. Similar conclusions to those in face recognition could be made: $l_1$-norm sparse regularization is useful in the phase of DL, and consequently in the phase of classification the $l_1$-regularized classifier is more powerful than the $l_2$-regularized one but with more computational cost.

### *6.1.4 The intra-class variance term in classification*

In the proposed GC and LC, there is an intra-class variance term defined on the solved representation vector, i.e., $w\cdot\left\|\hat{\boldsymbol{\alpha}}-\boldsymbol{m}_i\right\|_2^2$ in Eq. (18) and $\gamma_2\left\|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{m}_i^i\right\|_2^2$ in Eq. (20). Let's evaluate if the introduction of this intra-class variance term can help to improve the final classification rate. We test GC by face recognition on the AR face database (Martinez and Benavente 1998), and test LC by handwritten digit recognition on the USPS database (Hull, J.J. 1994) using 200 training samples per digit. (More information about the experiment settings can be found in Section 6.3 and Section 6.4, respectively.) By fixing $\gamma$ to 0.01 in GC and fixing $\gamma_1$ to 0.1 in LC, we report the classification rates by varying $w$ in GC and $\gamma_2$ in LC.

**Table 8:** The face recognition rates (%) by GC with different $w$ on the AR dataset.

| $w$ | 0 | 0.001 | 0.003 | 0.005 | 0.007 | 0.01 | 0.05 | 0.1 | 1 |
|---|---|---|---|---|---|---|---|---|---|
| FDDL | 91.6 | 92.0 | 92.1 | 92.3 | 92.4 | 92.4 | **92.6** | **92.6** | **92.6** |
| Simplified-FDDL | 91.1 | 91.4 | 91.6 | 91.8 | 92.0 | 92.0 | 92.0 | 92.0 | **92.1** |

**Table 9:** The handwritten digit recognition rates (%) by LC with different $\gamma_2$ on the USPS dataset.

| $\gamma_2$ | 0 | 0.001 | 0.002 | 0.003 | 0.005 | 0.01 |
|---|---|---|---|---|---|---|
| FDDL | 94.5 | 94.8 | 94.8 | 94.8 | 94.8 | **94.9** |
| Simplified-FDDL | 94.5 | 94.8 | 94.8 | **94.9** | **94.9** | **94.9** |

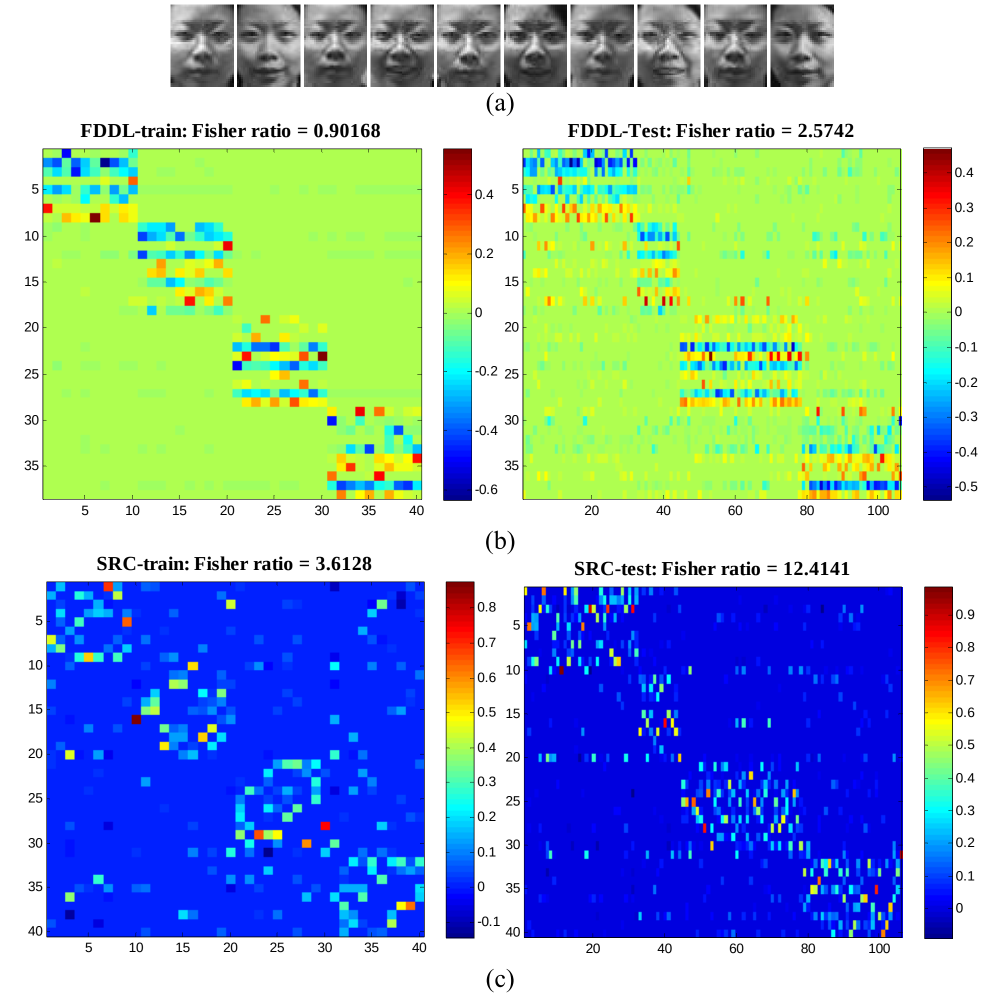
Table 8 shows the face recognition rates with different $w$. One can clearly see that the intra-class variance of $\hat{\boldsymbol{\alpha}}$ would benefit the final classification performance, which is in accordance with our analysis in Section 5. For example, GC with $w$=0.05 could have 1.0% improvement over GC with $w$=0 (i.e., without using the intra-class variance term). The digit recognition results by LC are reported in Table 9. Again, the intra-class variance term could bring certain benefit (about 0.4%) in classification. The benefit is not as obvious as that in GC because the sub-dictionary used in LC is much smaller than the whole dictionary used in GC so that the variation of representation coefficients in LC is generally smaller than that in GC, and hence the intra-class variance term in LC would not affect the final classification result as much as that in GC.

*6.1.5 Parameter selection by cross-validation*

There are four parameters need to be tuned in the proposed FDDL scheme, two in the DL model ($\lambda_1$ and $\lambda_2$) and two in the classifier ($\gamma$and $w$ in GC, or $\gamma_1$ and $\gamma_2$ in LC). In all the experiments, if no specific instructions, the tuning parameters in FDDL and the competing methods are evaluated by 5-fold cross validation. Based on our extensive experiment experience, the selection of $w$ (or $\gamma_2$) is relatively independent of the selection of other parameters. Therefore, to reduce the complexity of cross validation, we could tune $w$ (or $\gamma_2$) and the other three parameters separately. More specifically, we initially set $w$ (or $\gamma_2$) to 0 (other small values such as 0.001 could lead to similar results) to search the optimal values of $\lambda_1$, $\lambda_2$ and $\gamma$ (or $\gamma_1$), and then fix $\lambda_1$, $\lambda_2$ and $\gamma$ (or $\gamma_1$) to search for the optimal value of $w$ (or $\gamma_2$). In general, we search $\lambda_1$, $\lambda_2$ and $\gamma$ (or $\gamma_1$) from a small set {0.001, 0.005, 0.01, 0.05, 0.1}, and set the search range of $w$ and $\gamma_2$ to [0.001, 0.1].

**6.2. Fisher discrimination enhancement by FDDL**

FDDL aims to learn a dictionary to enhance the Fisher discrimination of representation coefficients. In this section, we evaluate if the Fisher discrimination criterion can be truly improved by using the learned dictionary ***D***. We first compare FDDL with SRC (Wright et al. 2009), which uses the original training samples as the dictionary. Four subjects in the FRGC dataset (Phillips et al. 2005) were randomly selected. Ten samples of each subject were used for training, and the remaining samples for testing. Fig. 4(a) shows the ten training samples of one subject; Fig. 4(b) illustrates the representation coefficient matrices of the training and test datasets by FDDL; and Fig. 4(c) illustrates the coefficient matrices of SRC. Please note that when we code a training sample by SRC, we take this sample away from the dictionary (i.e., using the leave-one-out strategy). One can see that by FDDL the coefficient matrix of the training dataset is nearly block diagonal, while each block is built by samples from the class corresponding to that sub-dictionary. In contrast, by SRC the coefficient matrix of the training dataset has many big non-block diagonal entries. For the test dataset, the coefficient matrix by FDDL is more regular than that by SRC. The Fisher ratio (i.e., $tr(\boldsymbol{S}_W(\boldsymbol{X}))/tr(\boldsymbol{S}_B(\boldsymbol{X}))$ of each coefficient matrix is computed and shown in Fig. 4. Clearly, the Fisher ratio values by FDDL are significantly lower than those by SRC on both the training and test datasets, validating the effectiveness of FDDL in enhancing the discrimination of representation coefficients.


(a)

FDDL-train: Fisher ratio = 0.90168

FDDL-Test: Fisher ratio = 2.5742

(b)

SRC-train: Fisher ratio = 3.6128

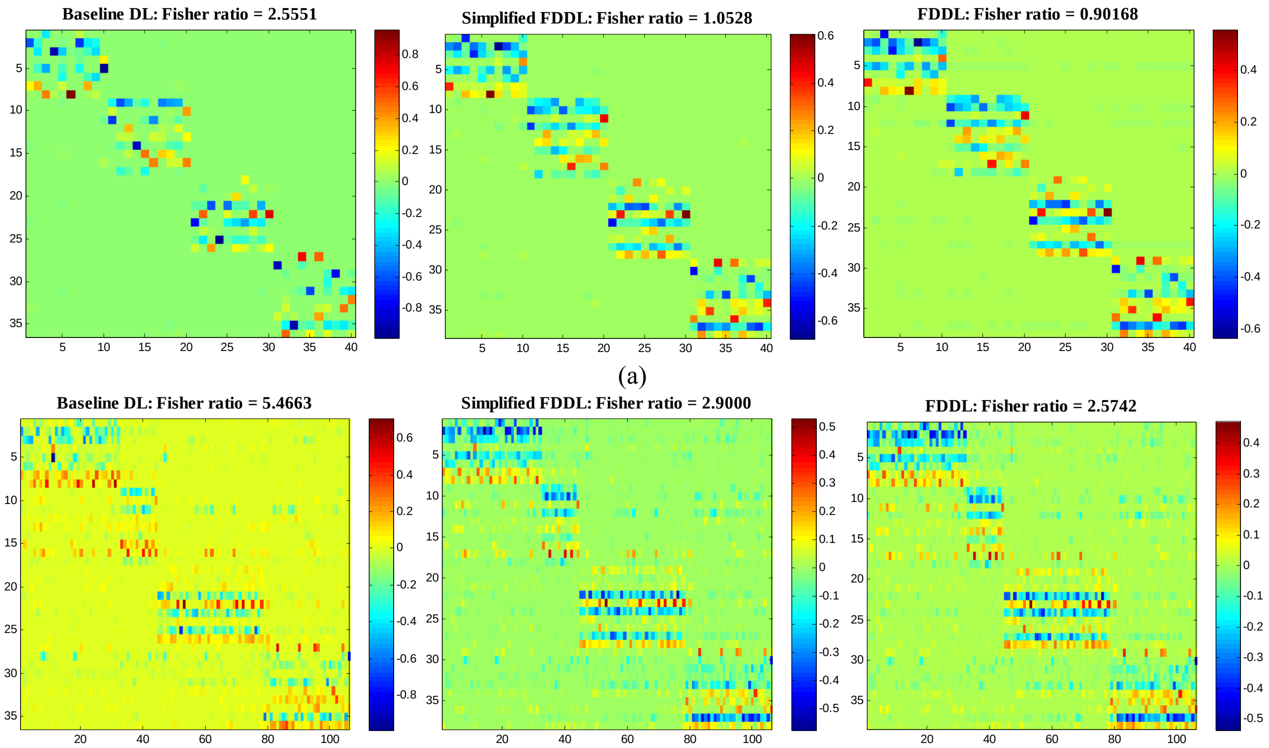
SRC-test: Fisher ratio = 12.4141

(c)

**Figure 4:** (a) The training samples from one subject. (b) The representation coefficient matrices by FDDL on the training (left) and test (right) datasets. (c) The representation coefficient matrices by SRC on the training (left) and test (right) datasets.

To more comprehensively evaluate the effectiveness of FDDL in improving the Fisher criterion, we further compare it with the baseline DL model in Eq. (3). The simplified FDDL model is also used in the comparison. The coefficient matrices by the three models on the training and test datasets are illustrated in Fig. 5(a) and Fig. 5(b), respectively. One can see that the baseline DL model will reduce the within-class scatter compared with SRC which does not learn a dictionary; however, its Fisher ratio is still much higher than simplified FDDL and FDDL.

The Fisher ratio by simplified FDDL is slightly higher than FDDL, showing that our simplification in the learning model does not sacrifice much the discrimination capability with much benefit in learning efficiency. Both the simplified FDDL model in Eq. (11) and the baseline DL model in Eq. (3) learn the class-specific sub-dictionaries class by class. However, compared with Eq. (3), Eq. (11) explicitly minimizes the within-class scatter of representation coefficients, which enhances much the discrimination of learned dictionary. This is why simplified FDDL has higher discrimination and better classification performance than the baseline DL.


Baseline DL: Fisher ratio = 2.5551

Simplified FDDL: Fisher ratio = 1.0528

FDDL: Fisher ratio = 0.90168

(a)

Baseline DL: Fisher ratio = 5.4663

Simplified FDDL: Fisher ratio = 2.9000

FDDL: Fisher ratio = 2.5742

(b)

**Figure 5:** The representation coefficient matrices by the baseline dictionary learning model (left), simplified FDDL (middle) and FDDL (right) on the (a) training dataset and (b) test dataset.

As discussed in Section 3.2, we employed the Fisher difference $tr(\boldsymbol{S}_W(\boldsymbol{X}))-a\cdot tr(\boldsymbol{S}_B(\boldsymbol{X}))$, instead of the Fisher ratio $tr(\boldsymbol{S}_W(\boldsymbol{X}))/tr(\boldsymbol{S}_B(\boldsymbol{X}))$, in the FDDL model (refer to Eq. (7)), and we set $a$=1 for simplicity. Let's evaluate if the setting of $a$ will affect much the final Fisher ratio value and the recognition rate. 100 subjects in the FRGC database (Phillips et al. 2005) are randomly selected in the evaluation. 10 images per subject are used as the training set, with the remaining as the test set (1,510 images in total). The images are cropped and normalized to

20×15. By fixing the other parameters in FDDL, Fig. 6 plots the Fisher ratio values and the recognition rates by setting $a$ to 0.1, 0.2, 0.4, 0.8, 1, 2, 4, 8 and 10, respectively. We can see that the resulting Fisher ratio drops slowly with the increase of $a$, and the final recognition rate is almost unchanged. Therefore, we set $a = 1$ in our FDDL model and it works very well in all our experiments.

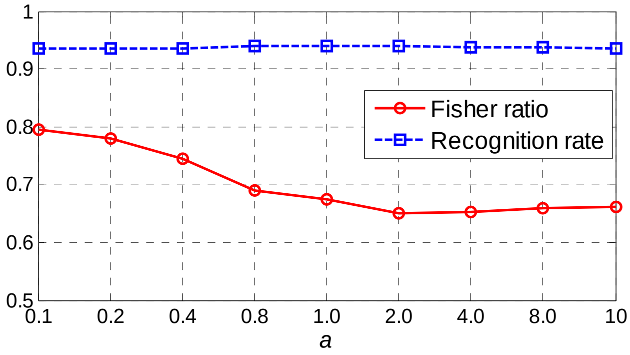


**Figure 6:** The Fisher ratio and recognition rate versus $a$.

### 6.3. Face recognition

We apply the proposed FDDL method to FR on the FRGC 2.0 (Phillips et al. 2005), AR (Martinez and Benavente 1998), and Multi-PIE (Gross et al. 2010) face databases. We compare FDDL with five latest DL based FR methods, including joint dictionary learning (JDL) (Zhou et al. 2012), dictionary learning with commonality and particularity (COPAR) (Kong and Wang 2012), label consistent KSVD (LCKSVD) (Jiang et al. 2013), discriminative KSVD (DKSVD) (Zhang and Li 2010) and dictionary learning with structure incoherence (DLSI) (Ramirez et al. 2010). We also compare with SRC (Wright et al. 2009) and two general classifiers, nearest subspace classifier (NSC) and linear support vector machine (SVM). Note that the original DLSI method and JDL method represent the query sample class by class. For a fair comparison, we also extended these two methods by representing the query sample on the whole dictionary and using the representation residual for classification (denoted by DLSI* and JDL*, respectively). The default number of dictionary atoms in FDDL is set as the number of training samples. The Eigenface feature (Turk and Pentland 1991) with dimension 300 is used in all FR experiments.

*a) FRGC database:* The FRGC version 2.0 (Phillips et al. 2005) is a large-scale face database established under uncontrolled indoor and outdoor settings. Some example images are shown in Fig. 7. We used a subset (316 subjects with no less than 10 samples, 7,318 images in total) of the query face dataset, which has large lighting, accessory (e.g., glasses), expression variations and image blur, etc. We randomly chose 2 to 5 samples per subject as the training set, and used the remaining images for testing. The images were cropped to 32×42 and all the experiments were run 10 times to calculate the mean and standard deviation. The results of FDDL, SRC, NSC, SVM, LCKSVD, DKSVD, JDL, COPAR, and DLSI are listed in Table 10. It can be seen that in most cases FDDL can have visible improvement over all the other methods. LCKSVD and DKSVD, which only use representation coefficients to do classification, do not work well. DLSI* and JDL* have better results than DLSI and JDL, respectively, which shows that representing the query image on the whole dictionary is more reasonable for FR tasks. COPAR underperforms FDDL by about 6%. This is mainly because FDDL employs the Fisher criterion to regularize the coding coefficients, which is more discriminative than the sparse regularization used in COPAR.

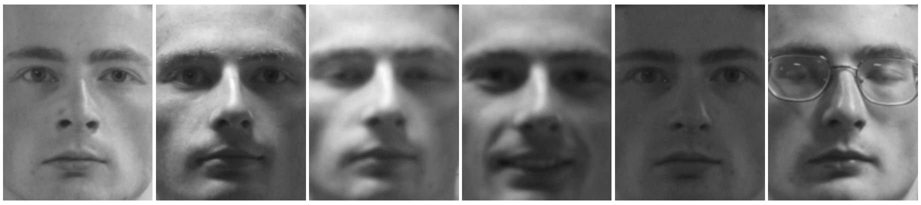


**Figure 7:** Some sample images from the FRGC 2.0 database

**Table 10:** The FR rates (%) of competing methods on the FRGC 2.0 database.

| $N_{ts}$ | SRC | NSC | SVM | DKSVD | LCKSVD | DLSI | DLSI* | COPAR | JDL | JDL* | **FDDL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 71.1±0.8 | 43.6±0.6 | 45.0±0.8 | 62.8±0.8 | 65.6±0.7 | 43.3±0.8 | 79.3±0.9 | 70.9±0.9 | 54.6±0.8 | 70.8±1.1 | **79.5±1.1** |
| 3 | 81.4±0.6 | 54.7±0.7 | 57.1±0.7 | 72.2±0.6 | 75.7±0.6 | 53.7±0.7 | 86.7±0.6 | 81.3±0.6 | 62.6±0.7 | 83.0±0.7 | **89.0±0.8** |
| 4 | 87.0±0.6 | 63.0±0.6 | 66.2±0.7 | 77.2±0.7 | 78.1±0.5 | 62.9±0.6 | 91.4±0.5 | 86.9±0.6 | 71.3±0.6 | 88.2±0.5 | **92.9±0.3** |
| 5 | 90.1±0.4 | 69.3±0.6 | 72.9±0.7 | 79.7±0.7 | 79.8±0.8 | 68.8±0.4 | 93.5±0.3 | 89.5±0.6 | 74.7±0.5 | 91.2±0.5 | **95.1±0.3** |

**Table 11:** The FR rates (%) of competing methods on the AR database.

| Method | SRC | NSC | SVM | DKSVD | LCKSVD | DLSI | DLSI* | COPAR | JDL | JDL* | **FDDL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 88.8 | 74.7 | 87.1 | 85.4 | 89.7 | 73.7 | 89.8 | 89.3 | 77.8 | 91.7 | **92.0** |

*b) FR on the AR database:* The cropped AR database (Martinez and Benavente 1998) consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separated sessions. As in (Wright et al. 2009), we chose a subset consisting of 50 male subjects and 50 female subjects in the experiment. For each subject, the 7 images with illumination and expression changes from Session 1 were used for training, and the other 7 images with the same condition from Session 2 were used for testing. The size of face image is 60×43. The recognition rates of FDDL and other competing methods are shown in Table 11. Again, we can see that FDDL has visible improvement over most of the competing methods. JDL* performs the second best in this experiment, followed by DLSI* and LCKSVD.

**Table 12:** The FR rates (%) of competing methods on the partial Multi-PIE dataset.

| Method | SRC | NSC | SVM | DKSVD | LCKSVD | DLSI | DLSI* | COPAR | JDL | JDL* | **FDDL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Test 1 | 95.5 | 90.8 | 91.6 | 93.9 | 93.7 | 91.4 | 94.1 | 95.3 | 90.0 | 96.1 | **96.7** |
| Test 2 | 96.1 | 94.7 | 92.2 | 89.8 | 90.8 | 94.9 | 95.9 | 96.3 | 91.0 | 96.3 | **98.0** |

**Table 13:** Recognition error rates (%) of competing methods on the whole Multi-PIE dataset.

| Method | U-SC | S-SC | MRR | MRR-LBP | RASR | RASR-IW | **FDDL** |
|---|---|---|---|---|---|---|---|
| Session2 | 5.4 | 4.8 | 6.3 | 4.7 | 6.1 | 5.0 | **4.3** |
| Session3 | 9.0 | 6.6 | 7.2 | 4.4 | 6.2 | 3.7 | **3.4** |
| Session4 | 7.5 | 4.9 | 7.0 | 4.4 | 7.7 | **2.7** | 3.1 |

*c) FR on the Multi-PIE database:* The CMU Multi-PIE face database (Gross et al. 2010) is a large scale database of 337 subjects including four sessions with simultaneous variations of pose, expression and illumination. We used the first 60 subjects presented in Session 1 as the training set. For each of the 60 training subjects, we used the frontal images of 14 illuminations[1], taken with neutral expression (for Test 1) or smile expression (for Test 2), for training. For the test set, we used the frontal images of 10 illuminations[2] from Session 3 with neutral expression (for Test 1) or smile expression (for Test 2). Note that Session 1 and Session 3 were recorded with a long time interval. The images were manually cropped and normalized to 100×82. For FDDL, the dictionary size of each class is set as half of the number of training samples. The experimental results of competing methods are listed in Table 12. We see that in both Test 1 and Test 2, FDDL works the best, followed by JDL*.

Through the above experiments on FRGC 2.0, AR and Multi-PIE databases, it is observed that DLSI* outperforms DLSI and JDL* outperforms JDL. This shows that representing the query sample on the whole dictionary is more effective than representing it on each class-specific sub-dictionary in the application of FR. DKSVD and LCKSVD are worse than FDDL, SRC, COPAR, DLSI* and JDL*, which implies that the representation residual is more powerful than the representation coefficients in face recognition. Meanwhile, FDDL outperforms COPAR and JDL, which again demonstrates that the utilization of both discriminative

[1] Illuminations {0,1,3,4, 6,7,8,11,13,14,16,17,18,19}.

[2] Illuminations {0,2,4,6,8,10,12,14,16,18}.

representation term and discriminative coefficient term could learn a more powerful dictionary.

In the above experiments on Multi-PIE, the frontal face images of the first 60 subjects were used in the training and testing, which is to show the advantages of FDDL over existing DL methods. To more comprehensively evaluate the performance of FDDL, we further conducted FR experiments on Multi-PIE using all the subjects. We adopted the experiment setting in supervised/unsupervised sparse coding (S-SC/U-SC) (Yang et al. 2010), misalignment-robust representation (MRR) (Yang et al. 2012), and robust alignment and illumination by sparse representation (RASR) (Wagner et al. 2012). We compare FDDL with these state-of-the-art sparse representation based methods which can deal with spatial transformation (e.g., misalignment) to some extent. All the 249 subjects presented in Session 1 are used in training. For each subject in the training set, the 7 frontal face images with neutral expression and extreme illuminations {0, 1, 7, 13, 14, 16, 18} are used. For the test set, all the face images with 20 illuminations, which are detected by the Viola and Jones' face detector (Viola & Jones, 2004), from Sessions 2-4 are used. In order to fairly compare with S-SC, U-SC, RASR and MRR, which use local patch features and/or involve misalignment correction, we performed face alignment by the method in MRR (Yang et al. 2012) and employed the 600-dimensional LBP histogram feature in FDDL. We also reported the results of MRR by using the 600-dimensioanl LBP histogram feature (MRR+LBP) and RASR with improve window (RASR-IW, Wagner et al. 2012). Table 13 lists the recognition rates of these competing methods. We can see that on Sessions 2-3, the proposed FDDL achieves better performance than others. On Session 4, FDDL is only slightly worse than RASR-IW.

### 6.4. Handwritten digit recognition

We then perform handwritten digit recognition on the widely used USPS database (Hull, J.J. 1994), which has 7,291 training and 2,007 test images. As in task-driven dictionary learning (TDDL, Mairal et al. 2012), we also artificially augmented the training set by shifting the digit images by 1 pixel in every direction. We compare FDDL with TDDL, COPAR, JDL and the handwritten digit recognition methods reported in (Huang and Aviyente 2006; Mairal et al. 2009; Ramirez et al. 2010). These methods include the state-of-the-art reconstructive DL methods with linear and bilinear classifier models (denoted by REC-L and REC-BL) (Mairal et al. 2009), the state-of-the-art supervised DL methods with generative training and discriminative training (denoted by SDL-G and SDL-D, Mairal et al. 2009), the state-of-the-art methods of sparse representation for signal classification (denoted by SRSC, Huang and Aviyente 2006) and DLSI (Ramirez et al. 2010). In addition, the results of some problem-specific methods (i.e., the standard Euclidean KNN and SVM with a Gaussian kernel) reported in (Ramirez et al. 2010) are also listed. Here the number of atoms in each sub-dictionary of FDDL is set to 200 and $\lambda_1=\gamma_1=0.1$, $\lambda_2=\gamma_2=0.001$.

Fig. 8 illustrates the learned dictionary atoms of digits 7 and 8, and Table 14 lists the recognition error rates of FDDL and the competing methods. We see that FDDL outperforms all the competing methods except for TDDL, while the recognition error rate of FDDL (2.89%) is very close to that of TDDL (2.84%). Meanwhile, it should be noted that TDDL learns a dictionary as well as an SVM classifier per class, and it performs classification with a one-versus-all strategy. In comparison, FDDL only learns a dictionary for each class, and its classifier (i.e., the LC presented in Section 5) is much simpler.

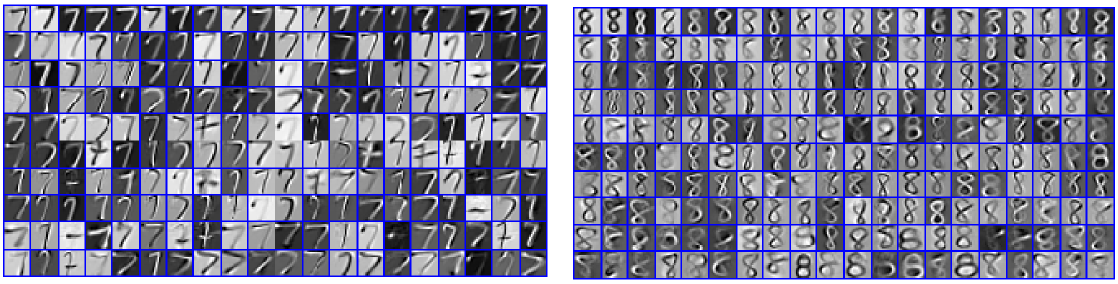


**Figure 8**: The learned atoms of digits 8 and 9 by FDDL.

**Table 14:** Recognition error rates (%) of competing methods on the USPS dataset.

| Algorithms | SRC | REC-L(BL) | SDL-G(D) | DLSI | KNN | SVM | TDDL | COPAR | JDL | **FDDL** |
|---|---|---|---|---|---|---|---|---|---|---|
| Error rate | 6.05 | 6.83(4.38) | 6.67(3.54) | 3.98 | 5.2 | 4.2 | **2.84** | 3.61 | 6.08 | 2.89 |

**Table 15:** The gender classification rates (%) of competing methods on the AR database.

| SRC | DKSVD | LCKSVD | DLSI | COPAR | JDL | SVM | NSC | **FDDL(LC)** | **FDDL(GC)** |
|---|---|---|---|---|---|---|---|---|---|
| 93.0 | 86.1 | 86.8 | 94.0 | 93.4 | 92.6 | 92.4 | 93.8 | **95.4** | 94.1 |

**Table 16:** The gender classification rates (%) of competing methods on the FRGC 2.0 database.

| SRC | DKSVD | LCKSVD | DLSI | COPAR | JDL | SVM | CNN | U-SC | S-SC | NSC | **FDDL(LC)** | **FDDL(GC)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 94.0 | 85.6 | 89.5 | 94.5 | 93.4 | 90.8 | 91.4 | 94.1 | 93.2 | 94.7 | 90.5 | **96.0** | 94.5 |

### 6.5. Gender classification

We first chose a non-occluded subset (14 images per subject) from the AR face database, which consists of 50 males and 50 females, to conduct experiments of gender classification. Images of the first 25 males and 25 females were used for training, and the remaining 25 males and 25 females were used for testing. PCA was used to reduce the dimension of each image to 300. Here 250 atoms per sub-dictionary are learned in FDDL. Table 15 lists the gender classification rates of the competing methods. It can be seen that FDDL with LC works the best with 1.4% improvement over the third best one, DLSI. FDDL with GC works the second best but it is 1.3% lower than FDDL with LC. This is because in gender classification, there are only two classes and each class has enough training samples so that the learned dictionary of each class is representative enough to represent the test sample.

We then evaluated FDDL on the large scale FRGC 2.0 database with the same experiment setting as that in (Yu et al. 2008) and (Yang et al. 2010). There are 568 individuals (243 females and 325 males) and 14,714 face images collected under various lighting conditions and backgrounds. We used the 3,014 images from randomly selected 114 subjects as the test set, and the rest 11,700 images as the training set. The 300-dimensional PCA feature is used in FDDL. The experimental results are listed in Table 16, where the state-of-the-art S-SC/U-SC methods in (Yang et al. 2010) and the CNN method in (Yu et al. 2008) are also reported. One can see that FDDL with LC outperforms all the competing methods, including those DL based ones (e.g., DLSI, LCKSVD, COPAR, JDL and SSC) and non-DL based ones (e.g., CNN).

### 6.6. Object categorization

Let’s then validate the effectiveness of FDDL on multi-class object categorization. The Oxford Flowers datasets with 17 categories (Nilsback and Zisserman 2006) is used. Some sample images are shown in Fig. 9. We adopted the default experiment settings provided on the website (www.robots.ox.ac.uk/~vgg/data/flowers), including the training, validation, test splits and the multiple features. It should be noted that these features are extracted from the flower regions which are well cropped by the preprocessing of segmentation.

**Figure 9:** Sample images of ‘daffodil’ from the Oxford Flowers dataset.

**Table 17:** The categorization accuracy (mean±std %) with single feature on the 17 category Oxford Flowers dataset.

| Features | NSC | SVM | MTJSRC-CG | SRC | COPAR | JDL | **FDDL** |
|---|---|---|---|---|---|---|---|
| | | (Gehler and Nowozin 2009) | (Yuan and Yan 2010) | | | | |
| Color | 61.7±3.3 | 60.9±2.1 | 64.0±3.3 | 61.9±2.2 | 61.1±4.0 | 58.6±6.1 | **65.0±2.4** |
| Shape | 69.9±3.2 | 70.3±1.3 | 71.5±0.8 | 72.7±1.9 | 72.5±1.2 | 63.8±1.6 | **72.8±1.7** |
| Texture | 55.8±1.4 | 63.7±2.7 | **67.6±2.2** | 61.4±0.9 | 60.4±0.7 | 49.1±0.9 | 64.9±1.7 |
| HSV | 61.3±0.7 | 62.9±2.3 | 65.0±3.9 | 62.5±3.0 | 62.7±2.5 | 61.7±0.3 | **65.5±3.4** |
| HOG | 57.4±3.0 | 58.5±4.5 | 62.6±2.7 | 61.4±1.9 | 61.4±2.2 | 51.4±2.2 | **62.7±2.4** |
| SIFTint | 70.7±0.7 | 70.6±1.6 | 74.0±2.2 | 73.7±2.9 | 74.1±3.3 | 64.8±3.4 | **74.4±2.6** |
| SIFTbdy | 61.9±4.2 | 59.4±3.3 | 63.2±3.3 | 62.3±2.6 | 62.6±1.8 | 48.4±0.2 | **64.0±2.4** |
| FLH | 79.3±4.3 | 88.6±1.7 | 88.4±2.7 | 88.4±2.7 | 88.6±1.4 | 90.1±2.2 | **91.7±1.2** |

**Table 18:** The categorization accuracy (mean±std %) with combined feature on the 17 category Oxford Flowers dataset.

| Related methods | Accuracy (%) |
|---|---|
| SRC combination | 85.9±2.2 |
| MKL (Gehler and Nowozin 2009) | 85.2±1.5 |
| LP-Boost (Gehler and Nowozin 2009) | 85.4±2.4 |
| CG-Boost (Gehler and Nowozin 2009) | 84.8±2.2 |
| COPAR | 85.9±0.8 |
| JDL | 81.7±2.0 |
| **FDDL** | 86.7±1.3 |
| MTJSRC-CG | 87.5±1.5 |
| **FDDL+MTJSRC** | **87.7±1.9** |
| Other state-of-the-art methods | |
| FLH+BOW (Fernando et al. 2012) | 94.5±1.5 |
| GRLF (Ye et al. 2012) | 91.7±1.7 |
| L1-BRD (Xie et al. 2010) | 89.0±0.6 |
| **FDDL with FLH feature** | **97.8±0.7** |

For a fair comparison with state-of-the-art methods such as MTJSRC (Yuan and Yan 2010), we also extended the original features from (Nilsback and Zisserman 2006; Nilsback and Zisserman 2008) to their kernel versions in the experiments. Specifically, we adopted the so-called column generation method (Yuan and Yan 2010). The idea is to generate a new descriptor for each original feature vector, and the new descriptor is composed of the similarities (e.g., the inner products) between this vector and all the training vectors in a higher dimensional kernel space. Given the original training dataset $\boldsymbol{A}$ and a test sample $\boldsymbol{y}$, their higher dimensional feature vectors can be written as $\phi(\boldsymbol{A})$=[ $\phi(\boldsymbol{a}_1)$,…, $\phi(\boldsymbol{a}_n)$] and $\phi(\boldsymbol{y})$, respectively, where $\boldsymbol{a}_i$ is the $i^{\text{th}}$ training sample and $\phi$ is the mapping function of the kernel. The similarities between $\boldsymbol{y}$ and all the vectors in $\boldsymbol{A}$ in the kernel space can be computed as $\boldsymbol{h}=\phi(\boldsymbol{A})^T\phi(\boldsymbol{y})$, and $\boldsymbol{h}$ is called the column-generation feature vector of $\boldsymbol{y}$. Similarly, the column-generation matrix of $\boldsymbol{A}$ is $\boldsymbol{G}=\phi(\boldsymbol{A})^T\phi(\boldsymbol{A})$. Here the kernel function is $\phi(\boldsymbol{a})^T\phi(\boldsymbol{y})=\exp\left(-\Omega(\boldsymbol{a},\boldsymbol{y})/\mu\right)$, where $\mu$ is set to the mean value of the pairwise Chi-square distances, denoted by Ω, on the training set. Finally, $\boldsymbol{G}$ and $\boldsymbol{h}$ take the place of $\boldsymbol{A}$ and $\boldsymbol{y}$, respectively, in the FDDL learning and testing.

The 17-category flower dataset consists of 17 species of flowers with 80 images per class. As in (Yuan and Yan 2010), we used the $\chi^2$ distance matrices of seven features (i.e., HSV, HOG, SIFTint, SIFTbdy, color, shape and texture vocabularies) to generate the training matrix $\boldsymbol{G}$ and test sample $\boldsymbol{h}$. We also used the histogram intersection similarity of the recently proposed Frequent Local Histogram (FLH, Fernando et al. 2012) feature to generate $\boldsymbol{G}$ and $\boldsymbol{h}$ in the experiment. Table 17 lists the best results of NSC, SVM, MTJSRC-CG, SRC, COPAR, JDL and the proposed FDDL on each single feature. Clearly, FDDL could always improve the original SRC (which directly uses training samples as the dictionary) by learning a discriminative dictionary, and it performs better than the two recently developed DL methods, COPAR and JDL. Compared to the other methods such as SVM and MTJSRC, FDDL achieves higher categorization rates in most cases.

We then evaluated the performance of FDDL by combining the seven features, and compared it with the corresponding state-of-the-art methods. The results are shown in Table 18. In order to more fairly compare with MTJSRC which is based on multi-task joint sparse representation, we also gave the results of FDDL+MTJSRC (i.e., all kinds of features are jointly represented on the dictionary learned by FDDL, and the joint representation error is used for classification). Here the parameters in learning dictionary are $\lambda_1$=0.01, $\lambda_2$=0.005, and all the task weights in MTJSRC are set to 1. By combining the seven features, MTJSRC, FDDL and FDDL+MTJSRC could achieve over 86.5% categorization rates, higher than the other competing methods. FDDL is slightly worse than MTJSRC but FDDL+MTJSRC is better than MTJSRC, which shows the effectiveness of FDDL in learning discriminative dictionaries. We further employed the FLH feature in FDDL and compared it with the latest state-of-the-arts on this flower dataset, including FLH+BOW (Fernando et al. 2012), L1-BRD (Xie et al. 2010) and GRLF (Ye et al. 2012). We see that the proposed FDDL achieves a categorization accuracy of 97.8%, which is the best result we can find on the Oxford 17-category flower dataset.

### 6.7. Action recognition

At last, we conduct action recognition on the UCF sport action dataset (Rodriguez et al. 2008) and the large scale UCF50 dataset (http://server.cs.ucf.edu/~vision/data.html). The video clips in the UCF sport action dataset were collected from various broadcast sports channels (e.g., BBC and ESPN). There are 140 videos in total and their action bank features can be found in (Sadanand et al. 2012). The videos cover 10 sport action classes: driving, golfing, kicking, lifting, horse riding, running, skateboarding, swinging-(prommel horse and floor), swinging-(high bar) and walking. The UCF50 dataset has 50 action categories (such as baseball pitch, biking, driving, skiing, and so on) and there are 6,680 realistic videos collected from YouTube.

On the UCF sport action dataset, we followed the experiment settings in (Qiu et al. 2011, Yao et al. 2010, and Jiang et al. 2013) and evaluated FDDL via five-fold cross validation, where one fold is used for testing and the remaining four folds for training. The action bank features (Sadanand et al. 2012) are used. We compare FDDL with SRC, KSVD, DKSVD, LC-KSVD (Jiang et al. 2013), COPAR, JDL and the methods in (Qiu et al. 2011, Yao et al. 2010, and Sadanand et al. 2012). The recognition rates are listed in Table 19. Clearly, FDDL shows better performance than all the other competing methods. In addition, by using the leave-one-video-out experiment setting in (Sadanand et al. 2012), the recognition accuracy of FDDL is 95.7%, while the accuracy of (Sadanand et al. 2012) is 95.0%.

Following the experiment settings in (Sadanand et al. 2012), we then evaluated FDDL on the large-scale UCF50 action dataset by using 5-fold group-wise cross validation, and compared it with the DL methods and the other state-of-the-art methods, including Oliva and Torralba 2001, Wang et al. 2009, and Sadanand et al. 2012. The results are shown in Table 20. Again, FDDL achieves better performance than all the competing methods. Compared with (Sadanand et al. 2012), FDDL has over 3% improvement.

**Table 19:** Recognition rates (%) on the UCF sports action dataset.

| Qiu et al. 2011 | Yao et al. 2010 | Sadanand et al. 2012 | SRC | KSVD | DKSVD | LCKSVD | COPAR | JDL | **FDDL** |
|---|---|---|---|---|---|---|---|---|---|
| 83.6 | 86.6 | 90.7 | 92.9 | 86.8 | 88.1 | 91.2 | 90.7 | 90.0 | **94.3** |

**Table 20:** Recognition rates (%) on the large-scale UCF50 action dataset.

| Oliva et al. 2001 | Wang et al. 2009 | Sadanand et al. 2012 | SRC | DKSVD | LCKSVD | COPAR | JDL | **FDDL** |
|---|---|---|---|---|---|---|---|---|
| 38.8 | 47.9 | 57.9 | 59.6 | 38.6 | 53.6 | 52.5 | 53.5 | **61.1** |

## 7. Conclusion

We proposed a sparse representation based Fisher Discrimination Dictionary Learning (FDDL) approach to image classification. The FDDL learns a structured dictionary whose sub-dictionaries have specific class labels. The discrimination capability of FDDL is two-folds. First, each sub-dictionary is trained to have good representation power to the samples from the corresponding class, but have poor representation power to the samples from other classes. Second, FDDL will result in discriminative coefficients by minimizing the with-class scatter and maximizing the between-class scatter of them. Consequently, we presented the classification schemes associated with FDDL, which use both the discriminative reconstruction residual and representation coefficients to classify the input query image. Extensive experimental results on face recognition, handwritten digit recognition, gender classification, object categorization and action recognition demonstrated the generality of FDDL and its superiority to many state-of-the-art dictionary learning based methods.

## 8. References

Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, *54*(1):4311–4322.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM. J. Imaging Science*, 2(1):183-202.

Bengio, S., Pereira, F., Singer, Y., & Strelow, D. (2009). Group sparse coding. In *Proc. Neural Information Processing Systems*.

Bobin, J., Starck, J., Fadili, J., Moudden, Y., & Donoho, D. (2007). Morphological component analysis: an adaptive thresholding strategy. *IEEE Trans. Image Processing*, *16*(11):2675-2681.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bryt, O. & Elad, M. (2008). Compression of facial images using the K-SVD algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282.

Candes, E. (2006). Compressive sampling. *Int. Congress of Mathematics*, *3*:1433–1452.

Castrodad, A., & Sapiro, G. (2012). Sparse modeling of human actions from motion imagery. *Int'l Journal of Computer Vision*, *100*:1-15.

Chai, Y., Lempitsky, V., & Zisserman, A. (2011). Bicos: A bi-level co-segmentation method for image classification. In *Proc. Int. Conf. Computer Vision*.

Cooley, J.W., & Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, *19*:297-301.

Deng, W.H., Hu, J.N., & Guo J. (2012). Extended SRC: Undersampled face recognition via intraclass variation dictionary, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(9): 1864-1870.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification (2nd ed.)*, Wiley-Interscience.

Elad, M. & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing*, *15*:(12):3736–3745.

Engan, K., Aase, S.O., & Husoy, J.H. (1999). Method of optimal directions for frame design. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*

Fernando, B., Fromont, E., & Tuytelaars, T. (2012). Effective use of frequent itemset mining for image classification. In *Proc. European Conf. Computer Vision*.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):643–660.

Gehler, P., & Nowozin, S. (2009). On feature combination for multiclass object classification. In *Proc. Int'l Conf. Computer Vision*.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Image and Vision Computing*, *28*:807–813.

Guha, T., & Ward, R.K. (2012). Learning Sparse Representations for Human Action Recognition. *IEEE Trans. Pattern Analysis and Machine Learning*, 34(8):1576-1888.

Guo, Y., Li, S., Yang, J., Shu, T., & Wu, L. (2003). A generalized Foley-Sammon transform based on generalized fisher discrimination criterion and its application to face recognition. *Pattern Recognition Letter*, 24(1-3): 147:158.

Hoyer, P.O. (2002). Non-negative sparse coding. In *Proc. IEEE Workshop Neural Networks for Signal Processing*.

Huang, K., & Aviyente, S. (2006). Sparse representation for signal classification. In *Proc. Neural Information and Processing Systems*.

Hull, J.J. (1994). A database for handwritten text recognition research. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(5):550–554.

Jenatton, R., Mairal, J., Obozinski, G., & Bach, F. (2011). Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297-2234.

Jia, Y.Q., Nie, F.P., & Zhang C.S. (2009). Trace ratio problem revisited. *IEEE Trans. Neural Netw.*, 20(4): 729-735.

Jiang, Z.L., Lin, Z., & Davis, L.S. (2013). Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, preprint.

Jiang, Z.L., Zhang, G.X., & Davis, L.S. (2012). Submodular Dictionary Learning for Sparse Coding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Kim, S.J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). A interior-point method for large-scale $l_1$-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing* 1, 606–617.

Kong, S., & Wang, D.H. (2012). A dictionary learning approach for classification: Separating the particularity and the commonality. In *Proc. European Conf. Computer Vision*.

Li, H., Jiang, T., & Zhang, K. (2006). Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans. Neural Netw.*, 17(1): 157-165.

Lian, X.C., Li, Z.W., Lu, B.L., & Zhang, L. (2010). Max-Margin Dictionary Learning for Multi-class Image Categorization. In *Proc. European Conf. Computer Vision*.

Mairal, J., Elad, M., & Sapiro, G. (2008). Sparse representation for color image restoration. *IEEE Trans. Image Processing*, *17*(1): 53–69.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zissserman, A. (2008). Learning discriminative dictionaries for local image analysis. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Mairal, J., Leordeanu, M., Bach, F., Hebert, M., & Ponce, J. (2008). Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation. In *Proc. European Conf. Computer Vision*.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Supervised dictionary learning. In *Proc. Neural Information and Processing Systems*.

Mairal, J., Bach, F., & Ponce, J. (2012). Task-Driven Dictionary Learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(4):791-804.

Mallat, S. (1999). *A wavelet Tour of Signal Processing*, second ed. Academic Press.

Martinez A., & Benavente, R. (1998). *The AR face database*. CVC Tech. Report No. 24.

Nesterov, Y. & Nemirovskii, A. (1994). Interior-point polynomial algorithms in convex programming, SIAM Philadelphia, PA.

Nilsback, M., & Zisserman, A. (2006). A visual vocabulary for flower classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145-174.

Okatani, T., & Deguchi, K. (2007). On the Wiberg algorithm for matrix factorization in the presence of missing components. *Int'l Journal of Computer Vision*, 72(3):329-337.

Olshausen, B.A., & Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*: 607-609.

Olshausen, B.A., & Field, D.J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research, 37*(23):3311-3325.

Pham, D., & Venkatesh, S. (2008). Joint learning and dictionary construction for pattern recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Phillips, P.J., Flynn, P.J., Scruggs, W.T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W.J. (2005). Overview of the face recognition grand challenge. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Qi, X.B., Xiao, R., Guo, J., & Zhang, L. (2012). Pairwise rotation invariant co-occurrence local binary pattern. In *Proc. European Conf. Computer Vision*.

Qiu, Q., Jiang, Z.L., & Chellappa, R. (2011). Sparse Dictionary-based Representation and Recognition of Action Attributes. In *Proc. Int'l Conf. Computer Vision*.

Ramirez, I., Sprechmann, P., & Sapiro, G. (2010). Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Rodriguez, F., & Sapiro, G. (2007). Sparse representation for image classification: Learning discriminative and reconstructive non-parametric dictionaries. *IMA Preprint 2213*.

Rodriguez, M., Ahmed, J., and Shah, M. (2008). A spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Rosasco, L., Verri, A., Santoro, M., Mosci, S., & Villa, S. (2009). *Iterative Projection Methods for Structured Sparsity Regularization.* MIT Technical Reports, MIT-CSAIL-TR-2009-050, CBCL-282.

Rubinstein, R., Bruckstein, A.M., & Elad, M. (2010). Dictionaries for Sparse Representation Modeling. In *Proceedings of the IEEE*, *98*(6):1045-1057.

Sadanand, S., & Corso, J.J. (2012). Action bank: A high-level representation of activeity in video. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Shen, L., Wang, S.H., Sun, G., Jiang, S.Q., & Huang, Q.M. (2013). Multi-level discriminative dictionary learning towards hierarchical visual categorization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Song, F.X., Zhang, D., Mei D.Y., & Guo, Z.W. (2007). A multiple maximum scatter difference discriminant criterion for facial feature extraction. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics, 37(6): 1599-1606.

Sprechmann, P., & Sapiro, G. (2010). Dictionary learning and sparse coding for unsupervised clustering. In *Proc. Int'l Conf. Acoustics Speech and Signal Processing*.

Szabo, Z., Poczos, B., & Lorincz, A. (2011). Online group-structured dictionary learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Tropp, J.A., & Wright, S.J. (2010). Computational methods for sparse solution of linear inverse problems. In *Proc. IEEE Conf. Special Issue on Applications of Compressive Sensing & Sparse Representation*, 98(6):948–958.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86.

Vinje, W.E., & Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *SCIENCE*, *287*(5456):1273-1276.

Viola, P., & Jones, M.J. (2004) Robust real-time face detection. *Int'l J. Computer Vision*, 57:137–154.

Wagner, A., Wright, J., Ganesh, A., Zhou, Z.H., Mobahi, H., & Ma, Y. (2012). Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(2): 373-386.

Wang, H., Yan, S.C., Xu, D., Tang, X.O., & Huang, T. (2007). Trace ratio vs. ratio trace for dimensionality reduction. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition.*

Wang, H., Ullah, M., Klaser, A., Laptev, I., & Schmid C. (2009). Evaluation of local spatio-temporal features for actions recognition. In *Proc. British Machine Vision Conference*.

Wang, H.R., Yuan, C.F., Hu, W.M., & Sun, C.Y. (2012). Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognition*, 45(11):3902-3911.

Wright, J.S., Nowak, D.R., & Figueiredo T.A.M. (2009a). Sparse reconstruction by separable approximation. *IEEE Trans. Signal Processing*, 57(7): 2479-2493.

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S, & Ma, Y. (2009). Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *31*(2):210–227.

Wu, Y.N., Si, Z.Z., Gong, H.F., & Zhu, S.C. (2010). Learning active basis model for object detection and recognition. *Int'l Journal of Computer Vision*, 90:198-235.

Xie, N., Ling, H., Hu, W., & Zhang, X. (2010). Use bin-ratio information for category and scene classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yang, A.Y., Ganesh, A., Zhou, Z.H., Sastry, S.S., & Ma, Y. (2010). A review of fast $l_1$-minimization algorithms for robust face recognition. *arXiv:1007.3753v2*.

Yang, J.C., Wright, J., Ma, Y., & Huang, T. (2008). Image super-resolution as sparse representation of raw image patches. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yang, J.C., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yang, J.C., Yu, K., & Huang, T. (2010). Supervised Translation-Invariant Sparse coding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yang, M. & Zhang, L., (2010). Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In *Proc. European Conf. Computer Vision*.

Yang, M., Zhang, L., Yang, J., & Zhang, D. (2010). Metaface learning for sparse representation based face recognition. In *Proc. IEEE Conf. Image Processing*.

Yang, M., Zhang, L., Yang, J., & Zhang, D. (2011). Robust sparse coding for face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yang, M., Zhang, L., Feng, X.C., & Zhang, D. (2011). Fisher Discrimination Dictionary Learning for sparse representation. In *Proc. Int'l Conf. Computer Vision*.

Yang, M., Zhang, L., & Zhang, D. (2012). Efficient misalignment robust representation for real-time face recognition. In *Proc. European Conf. Computer Vision*.

Yao, A., Gall, J., & Gool, L.V. (2010). A hough transform-based voting framework for action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Ye, G.N., Liu, D., Jhuo I-H., & Chang S-F. (2012). Robust late fusion with rank minimization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Yu, K., Xu, W., & Gong, Y. (2009). Deep learning with kernel regularization for visual recognition. In *In Advances in Neural Information Processing Systems* 21,.

Yuan, X.T., & Yan, S.C. (2010). Visual classification with multitask joint sparse representation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Zhang, L., Yang, M., & Feng, X.C. (2011). Sparse representation or collaborative representation: which helps face recognition? In *Proc. Int'l Conf. Computer Vision*.

Zhang, Q., & Li, B.X. (2010). Discriminative K-SVD for dictionary learning in face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Zhang, Z.D., Ganesh, A., Liang, X., & Ma, Y. (2012). TILT: Transformation invariant low-rank textures. *Int'l Journal of Computer Vision, 99*: 1-24.

Zhou, M.Y., Chen, H.J., Paisley, J., Ren, L., Li, L.B., Xing, Z.M., Dunson, D., Sapiro, G., & Carin, L. (2012). Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images. *IEEE Trans. Image Processing*, *21*(1):130-144.

Zhou, N., & Fan J.P. (2012). Learning inter-related visual dictionary for object recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via elastic net. *J. R. Statist. Soc. B*, 67, Part 2: 301-320.

## Appendix 1: *tr*($\boldsymbol{S}_B$(*X*)) when $X_i^j=0,\ j\neq i$

Denote by $\boldsymbol{m}_i^i$, $\boldsymbol{m}_i$ and $\boldsymbol{m}$ the mean vectors of $\boldsymbol{X}_i^i$, $\boldsymbol{X}_i$ and $\boldsymbol{X}$, respectively. Because $\boldsymbol{X}_i^j=0$ for $j\neq i$, we can rewrite $\boldsymbol{m}_i=\left[\boldsymbol{0};\cdots;\boldsymbol{m}_i^i;\cdots;\boldsymbol{0}\right]$ and $\boldsymbol{m}=\left[n_1\boldsymbol{m}_1^1;\cdots;n_i\boldsymbol{m}_i^i;\cdots;n_K\boldsymbol{m}_K^K\right]/n$ . Therefore, the between-class scatter, i.e., $tr\left(\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)=\sum_{i=1}^{K}n_i\left\|\boldsymbol{m}_i-\boldsymbol{m}\right\|_2^2$, becomes

$$\boldsymbol{S}_B\left(\boldsymbol{X}\right)=\sum_{i=1}^{K}n_i/n^2\left[-n_1\boldsymbol{m}_1^1;\cdots;\left(n-n_i\right)\boldsymbol{m}_i^i;\cdots;-n_K\boldsymbol{m}_K^K\right]\left[-n_1\boldsymbol{m}_1^1;\cdots;\left(n-n_i\right)\boldsymbol{m}_i^i;\cdots;-n_K\boldsymbol{m}_K^K\right]^T .$$

Denote by $\kappa_i$=1-$n_i/n$. After some derivation, the trace of $\boldsymbol{S}_B$($\boldsymbol{X}$) becomes

$$tr\left(\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)=\sum_{i=1}^{K}n_i/n^2\left\|\left[-n_1\boldsymbol{m}_1^1;\cdots;\left(n-n_i\right)\boldsymbol{m}_i^i;\cdots;-n_K\boldsymbol{m}_K^K\right]\right\|_2^2=\sum_{i=1}^{K}\kappa_i n_i\left\|\boldsymbol{m}_i^i\right\|_2^2 .$$

Because $\boldsymbol{m}_i^i$ is the mean representation vector of the samples from the same class, which will generally have non-neglected values, the trace of between-class scatter will have big energy in general.

## Appendix 2: The derivation of simplified FDDL model

Denote by $\boldsymbol{m}_i^i$ and $\boldsymbol{m}_i$ the mean vector of $\boldsymbol{X}_i^i$ and $\boldsymbol{X}_i$, respectively. Because $\boldsymbol{X}_i^j=0$ for $j\neq i$, we can rewrite $\boldsymbol{m}_i=\left[\boldsymbol{0};\cdots;\boldsymbol{m}_i^i;\cdots;\boldsymbol{0}\right]$. So the within-class scatter changes to

$$\boldsymbol{S}_W\left(\boldsymbol{X}\right)=\sum_{i=1}^{K}\sum_{\boldsymbol{x}_k\in X_i}\left(\boldsymbol{x}_k^i-\boldsymbol{m}_i^i\right)\left(\boldsymbol{x}_k^i-\boldsymbol{m}_i^i\right)^T .$$

The trace of within-class scatter is

$$tr\left(\boldsymbol{S}_W\left(\boldsymbol{X}\right)\right)=\sum_{i=1}^{K}\sum_{\boldsymbol{x}_k\in X_i}\left\|\boldsymbol{x}_k^i-\boldsymbol{m}_i^i\right\|_2^2 .$$

Based on **Appendix 1**, the trace of between-class scatter is $tr\left(\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)=\sum_{i=1}^{K}\kappa_i n_i\left\|\boldsymbol{m}_i^i\right\|_2^2$ , where $\kappa_i$=1-$n_i/n$. Therefore the discriminative coefficient term, i.e., $f\left(\boldsymbol{X}\right)=\left(\boldsymbol{S}_W\left(\boldsymbol{X}\right)-\boldsymbol{S}_B\left(\boldsymbol{X}\right)\right)+\eta\left\|\boldsymbol{X}\right\|_F^2$, could be simplified to

$$f\left(\boldsymbol{X}\right)=\sum_{i=1}^{K}\left(\sum_{\boldsymbol{x}_k\in X_i}\left\|\boldsymbol{x}_k^i-\boldsymbol{m}_i^i\right\|_2^2+\kappa_i\left(\left\|\boldsymbol{X}_i^i\right\|_F^2-n_i\left\|\boldsymbol{m}_i^i\right\|_2^2\right)+\left(\eta-\kappa_i\right)\left\|\boldsymbol{X}_i^i\right\|_F^2\right).$$

Denote by $\boldsymbol{E}_i^j=\left[1\right]_{n_i\times n_j}$ the matrix of size $n_i\times n_j$ with all entries being 1, then $\boldsymbol{M}_i^i=\left[\boldsymbol{m}_i^i\right]_{1\times n_i}=\boldsymbol{X}_i^i\boldsymbol{E}_i^i/n_i$ . Because ***I***- $\boldsymbol{E}_i^i/n_i\left(\boldsymbol{E}_i^i/n_i\right)^T$ = (***I***- $\boldsymbol{E}_i^i/n_i$ )(***I***- $\boldsymbol{E}_i^i/n_i$ )$^T$, we have

$$\begin{aligned}&\left\|\boldsymbol{X}_i^i\right\|_F^2 - n_i\left\|\boldsymbol{m}_i^i\right\|_2^2 = \left\|\boldsymbol{X}_i^i\right\|_F^2 - \left\|\left[\boldsymbol{m}_i^i\right]_{1\times n_i}\right\|_F^2 = tr\left(\boldsymbol{X}_i^i\left(\boldsymbol{I}-\boldsymbol{E}_i^i/n_i\left(\boldsymbol{E}_i^i/n_i\right)^T\right)\left(\boldsymbol{X}_i^i\right)^T\right)\\ &= tr\left(\boldsymbol{X}_i^i\left(\boldsymbol{I}-\boldsymbol{E}_i^i/n_i\right)\left(\boldsymbol{I}-\boldsymbol{E}_i^i/n_i\right)^T\left(\boldsymbol{X}_i^i\right)^T\right) = \left\|\boldsymbol{X}_i^i - \left[\boldsymbol{m}_i^i\right]_{1\times n_i}\right\|_F^2 = \left\|\boldsymbol{X}_i^i - \boldsymbol{M}_i^i\right\|_F^2\end{aligned}$$

Then the discriminative coefficient term could be written as

$$\begin{aligned}f(\boldsymbol{X}) &= \sum\nolimits_{i=1}^K\left(\sum\nolimits_{\boldsymbol{x}_k\in X_i}\left\|\boldsymbol{x}_k^i - \boldsymbol{m}_i^i\right\|_2^2 + \kappa_i\left\|\boldsymbol{X}_i^i - \boldsymbol{M}_i^i\right\|_F^2 + (\eta-\kappa_i)\left\|\boldsymbol{X}_i^i\right\|_F^2\right)\\ &= \sum\nolimits_{i=1}^K\left((1+\kappa_i)\left\|\boldsymbol{X}_i^i - \boldsymbol{M}_i^i\right\|_F^2 + (\eta-\kappa_i)\left\|\boldsymbol{X}_i^i\right\|_F^2\right)\end{aligned} \tag{21}$$

With the constraint that $\boldsymbol{X}_i^j = 0$ for $j \neq i$ in Eq. (10), we have

$$\left\|\boldsymbol{A}_i - \boldsymbol{D}\boldsymbol{X}_i\right\|_F^2 = \left\|\boldsymbol{A}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2 \tag{22}$$

With Eq. (21) and Eq. (22), the model of simplified FDDL (i.e., Eq. (10)) could be written as

$$\min_{\boldsymbol{D},\boldsymbol{X}}\sum\nolimits_{i=1}^K\left(\left\|\boldsymbol{A}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2 + \lambda_1'\left\|\boldsymbol{X}_i^i\right\|_1 + \lambda_2'\left\|\boldsymbol{X}_i^i - \boldsymbol{M}_i^i\right\|_F^2 + \lambda_3'\left\|\boldsymbol{X}_i^i\right\|_F^2\right) \text{ s.t. } \left\|\boldsymbol{d}_n\right\|_2 = 1, \forall n \tag{23}$$

where $\lambda_1' = \lambda_1/2$, $\lambda_2' = \lambda_2(1+\kappa_i)/2$, and $\lambda_3' = \lambda_2(\eta-\kappa_i)/2$.

## Appendix 3: The convexity of *f*$_i$(*X*)

Let $\boldsymbol{E}_i^j = [1]_{n_i\times n_j}$ be a matrix of size $n_i \times n_j$ with all entries being 1, and let $\boldsymbol{N}_i = \boldsymbol{I}_{n_i\times n_i} - \boldsymbol{E}_i^i/n_i$, $\boldsymbol{P}_i = \boldsymbol{E}_i^i/n_i - \boldsymbol{E}_i^i/n$, $\boldsymbol{C}_i^j = \boldsymbol{E}_i^j/n$, where $\boldsymbol{I}_{n_i\times n_i}$ is an identity matrix of size $n_i \times n_i$.

From $f_i(\boldsymbol{X}_i) = \left\|\boldsymbol{X}_i - \boldsymbol{M}_i\right\|_F^2 - \sum_{k=1}^K\left\|\boldsymbol{M}_k - \boldsymbol{M}\right\|_F^2 + \eta\left\|\boldsymbol{X}_i\right\|_F^2$, we can derive that

$$f_i(\boldsymbol{X}_i) = \left\|\boldsymbol{X}_i\boldsymbol{N}_i\right\|_F^2 - \left\|\boldsymbol{X}_i\boldsymbol{P}_i - \boldsymbol{G}\right\|_F^2 - \sum\nolimits_{k=1,k\neq i}^K\left\|\boldsymbol{Z}_k - \boldsymbol{X}_i\boldsymbol{C}_i^k\right\|_F^2 + \eta\left\|\boldsymbol{X}_i\right\|_F^2 \tag{24}$$

where $\boldsymbol{G} = \sum_{k=1,k\neq i}^K\boldsymbol{X}_k\boldsymbol{C}_k^i$, $\boldsymbol{Z}_k = \boldsymbol{X}_k\boldsymbol{E}_k^k/n_k - \sum_{j=1,j\neq i}^K\boldsymbol{X}_j\boldsymbol{C}_j^k$.

Rewrite $\boldsymbol{X}_i$ as a column vector, $\boldsymbol{\chi}_i = \left[\boldsymbol{r}_{i,1}, \boldsymbol{r}_{i,2}, \cdots, \boldsymbol{r}_{i,d}\right]^T$, where $\boldsymbol{r}_{i,j}$ is the $j^{\text{th}}$ row vector of $\boldsymbol{X}_i$, and $d$ is the total number of row vectors in $\boldsymbol{X}_i$. Then $f_i(\boldsymbol{X}_i)$ equals to

$$\left\|\text{diag}\left(\boldsymbol{N}_i^T\right)\boldsymbol{\chi}_i\right\|_2^2 - \left\|\text{diag}\left(\boldsymbol{P}_i^T\right)\boldsymbol{\chi}_i - \text{vec}\left(\boldsymbol{G}^T\right)\right\|_2^2 - \sum\nolimits_{k=1,k\neq i}^K\left\|\text{diag}\left(\left(\boldsymbol{C}_i^k\right)^T\right)\boldsymbol{\chi}_i - \text{vec}\left(\boldsymbol{Z}_k^T\right)\right\|_2^2 + \eta\left\|\boldsymbol{\chi}_i\right\|_2^2$$

where diag($\boldsymbol{T}$) is to construct a block diagonal matrix with each block on the diagonal being matrix $\boldsymbol{T}$, and vec($\boldsymbol{T}$) is to construct a column vector by concatenating all the column vectors of $\boldsymbol{T}$.

The convexity of $f_i(\boldsymbol{\chi}_i)$ depends on whether its Hessian matrix $\nabla^2 f_i(\boldsymbol{\chi}_i)$ is positive definite or not (Boyd and Vandenberghe 2004). We could write the Hessian matrix of $f_i(\boldsymbol{\chi}_i)$ as

$$\nabla^2 f_i\left(\boldsymbol{\chi}_i\right) = 2\text{diag}\left(\boldsymbol{N}_i\boldsymbol{N}_i^T\right) - 2\text{diag}\left(\boldsymbol{P}_i\boldsymbol{P}_i^T\right) - \sum\nolimits_{k=1,k\neq i}^{K} 2\text{diag}\left(\boldsymbol{C}_i^k\left(\boldsymbol{C}_i^k\right)^T\right) + 2\eta\boldsymbol{I}\,.$$

$\nabla^2 f_i(\boldsymbol{\chi}_i)$ will be positive definite if the following matrix $\boldsymbol{S}$ is positive definite:

$$\boldsymbol{S} = \boldsymbol{N}_i\boldsymbol{N}_i^T - \left(\boldsymbol{P}_i\boldsymbol{P}_i^T + \sum\nolimits_{k=1,k\neq i}^{K} \boldsymbol{C}_i^k\left(\boldsymbol{C}_i^k\right)^T\right) + \eta\boldsymbol{I}\ \ .$$

After some derivations, we have

$$\boldsymbol{S} = \left(1+\eta\right)\boldsymbol{I} - \boldsymbol{E}_i^i\left(2/n_i - 2/n + \sum\nolimits_{k=1}^{K} n_k/n^2\right).$$

In order to make $\boldsymbol{S}$ positive define, each eigenvalue of $\boldsymbol{S}$ should be greater than 0. Because the maximal eigenvalue of $\boldsymbol{E}_i^i$ is $n_i$, we should ensure

$$\left(1+\eta\right) - n_i\left(2/n_i - 2/n + \sum\nolimits_{k=1}^{K} n_k/n^2\right) > 0$$

For $n=n_1+n_2+\ldots+n_K$, we have $\eta>\kappa_i$, which could guarantee that $f_i(\boldsymbol{X}_i)$ is convex to $\boldsymbol{X}_i$. Here $\kappa_i=1-n_i/n$.