Fisher Discrimination Dictionary Learning for Sparse Representation

Meng Yang[a], Lei Zhang[a], Xiangchu Feng[b], and David Zhang[a]
[a]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b]Dept. of Applied Mathematics, Xidian University, Xi'an, China
{csmyang, cslzhang}@comp.polyu.edu.hk

Abstract

Sparse representation based classification has led to interesting image recognition results, while the dictionary used for sparse coding plays a key role in it. This paper presents a novel dictionary learning (DL) method to improve the pattern classification performance. Based on the Fisher discrimination criterion, a structured dictionary, whose dictionary atoms have correspondence to the class labels, is learned so that the reconstruction error after sparse coding can be used for pattern classification. Meanwhile, the Fisher discrimination criterion is imposed on the coding coefficients so that they have small within-class scatter but big between-class scatter. A new classification scheme associated with the proposed Fisher discrimination DL (FDDL) method is then presented by using both the discriminative information in the reconstruction error and sparse coding coefficients. The proposed FDDL is extensively evaluated on benchmark image databases in comparison with existing sparse representation and DL based classification methods.

1. Introduction

The past several years have witnessed the rapid development of the theory and algorithms of sparse representation (or coding) [30] and its successful applications in image restoration [1-3] and compressed sensing [4]. Recently sparse representation techniques have also led to promising results in image classification, e.g. face recognition (FR) [5-7, 10, 31], digit and texture classification [8-9, 11-12], etc. The success of sparse representation based classification owes to the fact that a high-dimensional image can be represented or coded by a few representative samples from the same class in a low-dimensional manifold, and the recent progress of l_0-norm and l_1-norm minimization techniques [28].

In sparse representation based classification, there are two phases: coding and classification. First, the query signal/image is collaboratively coded over a dictionary of atoms with some sparsity constraint, and then classification is performed based on the coding coefficients and the dictionary. The dictionary for sparse coding could be predefined. For example, Wright *et al.* [5] directly used the training samples of all classes as the dictionary to code the query face image, and classified the query face image by evaluating which class leads to the minimal reconstruction error. Although this so called sparse representation based classification (SRC) scheme shows interesting FR results, the dictionary used in it may not be effective enough to represent the query images due to the uncertain and noisy information in the original training images. The number of atoms of such a dictionary can also be very big, which increases the coding complexity. In addition, using the original training samples as the dictionary could not fully exploit the discriminative information hidden in the training samples. On the other hand, using analytically designed off-the-shelf bases as dictionary (e.g., [8] uses Haar wavelets and Gabor wavelets as the dictionary) might be universal to all types of images but will not be effective enough for specific type of images such as face, digit and texture images. In fact, all the above mentioned problems of predefined dictionary can be addressed, at least to some extent, by learning properly a non-parametric dictionary from the original training samples.

Dictionary learning (DL) aims to learn from the training samples the space where the given signal could be well represented or coded for processing. Many DL methods have been proposed for image processing [1, 3, 17] and classification [9-16]. One representative DL method for image processing is the KSVD algorithm [17], which learns an over-complete dictionary from a training dataset of natural image patches. However, KSVD is not suitable for classification tasks because it only requires that the learned dictionary could faithfully represent the training samples. Based on KSVD, Mairal *et al.* [14] added a discriminative reconstruction constraint in the DL model to gain discrimination ability, and used the learned dictionary for texture segmentation and scene analysis; however, this method is not convex and it does not explore the discrimination capability of sparse coding coefficients. Later, Mairal *et al.* [9] proposed a discriminative DL method by training a classifier of the coding coefficients, and verified their method for digit recognition and texture

classification. In [16], Pham *et al.* proposed a joint learning and dictionary construction method with consideration of the linear classifier performance and applied their method to object categorization and FR. Based on [16], Zhang *et al.* [10] proposed an algorithm called discriminative KSVD (DKSVD) for FR. All the works in [9], [10] and [16] try to learn a common dictionary shared by all classes, as well as a classifier of coefficients for classification. However, the shared dictionary loses the correspondence between the dictionary atoms and the class labels, and hence performing classification based on the reconstruction error associated with each class is not allowed. Different from these works, Yang *et al.* [13] learned a dictionary for each class and obtained better FR results than SRC. Ramirez *et al.* [11] used an incoherence promoting term to make the dictionaries associated with different classes as independent as possible. These methods use the reconstruction error associated with each class as the discriminative information for classification, but they do not enforce discriminative information into the sparse coding coefficients.

In this paper we propose a new discriminative DL framework which employs the Fisher discrimination criterion to learn a structured dictionary (i.e. the dictionary atoms have correspondence to the class labels so that the reconstruction error associated with each class can be used for classification). Meanwhile, the Fisher discrimination criterion is imposed on the coding coefficients to make them discriminative. To this end, in the DL process we make the sparse coding coefficients have small within-class scatter but big between-class scatter, and at the same time we make each class-specific sub-dictionary in the whole structured dictionary have good representation ability to the training samples from the associated class but poor representation ability for other classes. With the proposed Fisher discrimination based DL (FDDL) method, both the reconstruction error and the coding coefficient will be discriminative, and hence a new classification scheme is proposed to exploit such information. The FDDL method is applied to face, digit and gender recognition to evaluate its performance. Compared with the SRC method [5], not only higher classification accuracy is got by FDDL, but also a smaller dictionary can be learnt (e.g., on the Extended Yale B database, the learnt dictionary by FDDL with only 8 atoms per class could still achieve better FR performance than SRC with 20 atoms per class). Compared with other state-of-the-art methods, FDDL has competitive performance in various pattern recognition tasks.

The rest of this paper is organized as follows. Section 2 briefly introduces the SRC scheme in [5]. 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 experiments, and Section 7 concludes the paper.

2. Brief introduction of SRC

Wright *et al.* [5] proposed the sparse representation based classification (SRC) method for robust face recognition (FR). Suppose that we have c classes of subjects, and let $A = [A_1, A_2, \ldots, A_c]$ be the set of original training samples, where A_i is the sub-set of the training samples from class i. Denote by $\boldsymbol{y}$ a testing sample. The procedures of SRC are as follows.

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

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

where γ is a scalar constant.

ii.) Do classification via

$$\text{identity}(\boldsymbol{y}) = \arg\min_i \{e_i\}, \quad (2)$$

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

SRC use the reconstruction error e_i associated with each class to do FR. Impressive results were reported in [5].

3. Fisher discrimination dictionary learning

To improve the performance of previous DL methods, we propose here a novel Fisher discrimination based DL (FDDL) scheme. Instead of learning a shared dictionary to all classes, we learn a structured dictionary $D = [D_1, D_2, \ldots, D_c]$, where D_i is the class-specified sub-dictionary associated with class i, and c is the total number of classes. With such a D, we could use the reconstruction error for classification, as that in the SRC method [5].

Denote by $A=[A_1, A_2, \ldots, A_c]$ the set of training samples, where A_i is the sub-set of the training samples from class i. Denote by X the coding coefficient matrix of A over D, i.e. $A \approx DX$. We can write X as $X = [X_1, X_2, \ldots, X_c]$, where X_i is the sub-matrix containing the coding coefficients of A_i over D. Apart from requiring that D should have powerful reconstruction capability of A, we also require that D should have powerful discriminative capability of images in A. To this end, we propose the following FDDL model:

$$J_{(D,X)} = \arg\min_{(D,X)} \left\{ r(A, D, X) + \lambda_1 \left\| X \right\|_1 + \lambda_2 f(X) \right\}, \quad (3)$$

where $r(A,D,X)$ is the discriminative fidelity term; $\|X\|_1$ is the sparsity constraint; $f(X)$ is a discrimination constraint imposed on the coefficient matrix X; and λ_1 and λ_2 are scalar parameters. Next let's discuss the design of $r(A,D,X)$ and $f(X)$ based on the Fisher discrimination criterion.

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

We can write X_i, the representation of A_i over D, as $X_i = [X_i^1; \ldots; X_i^j; \ldots; X_i^c]$, where X_i^j is the coding coefficient of A_i over the sub-dictionary D_j. Denote the representation of D_k to A_i as $R_k = D_k X_i^k$. First of all, the dictionary D should be able to well represent A_i, and there is $A_i \approx DX_i = D_1 X_i^1 + \ldots +$

$D_1X_i^1+\ldots+D_cX_i^c=R_1+\ldots+R_i+\ldots+R_c$. Second, since D_i is associated with the i^{th} class, it is expected that A_i should be well represented by D_i but not by $D_j, j\neq i$. This implies that X_i^i should have some significant coefficients such that $\|A_i-D_iX_i^i\|_F^2$ is small, while X_i^j should have nearly zero coefficients such that $\|D_jX_i^j\|_F^2$ is small. Thus we define the discriminative fidelity term as

$$r(A_i,D,X_i)=\|A_i-DX_i\|_F^2+\|A_i-D_iX_i^i\|_F^2+\sum_{j=1,j\neq i}^{c}\|D_jX_i^j\|_F^2 . \quad (4)$$

An intuitive explanation of three terms in $r(A_i,D,X_i)$ is shown in Fig. 1. Fig. 1(a) shows that although D is ensured to represent A_i well, R_i may deviate much from A_i so that D_i could not well represent A_i. If we add another constraint that $\|A_i-D_iX_i^i\|_F^2$ is small, better discrimination will be achieved, as shown in Fig. 1(b). Nonetheless, A_i may also be well represented by other sub-dictionaries, e.g. D_{i-1} in Fig. 1(b), which reduces the discrimination capability of D. With the third constraint that the representation of $D_j, j\neq i$, to A_i is small, the proposed discriminative fidelity term could overcome this problem, as shown in Fig.1(c).

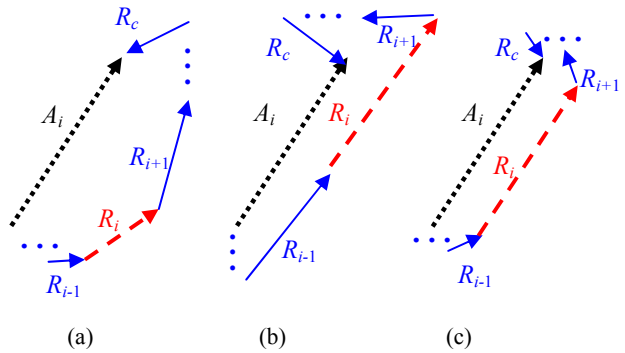

Figure 1: Illustration of the discriminative fidelity term. (a) Only D is required to well represent A_i. (b) Both D and D_i are required to well represent A_i. (c) The discriminative fidelity term in Eq. (4).

3.2. Discriminative coefficient term $f(X)$

To make dictionary D be discriminative for the samples in A, we can make the coding coefficient of A over D, i.e. X, be discriminative. Based on some criterion such as the Fisher discrimination criterion [18], this can be achieved by minimizing the within-class scatter of X, denoted by $S_W(X)$, and maximizing the between-class scatter of X, denoted by $S_B(X)$. $S_W(X)$ and $S_B(X)$ are defined as

$$S_W(X)=\sum_{i=1}^{c}\sum_{\boldsymbol{x}_k\in X_i}(\boldsymbol{x}_k-\boldsymbol{m}_i)(\boldsymbol{x}_k-\boldsymbol{m}_i)^T ,$$

$$S_B(X)=\sum_{i=1}^{c}n_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T ,$$

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

Intuitively, we can define $f(X)$ as $tr(S_W(X))-tr(S_B(X))$. However, such an $f(X)$ is non-convex and unstable. To solve this problem, we propose to add an elastic term $\|X\|_F^2$ into $f(X)$. So $f(X)$ is defined as

$$f(X)=tr(S_W(X))-tr(S_B(X))+\eta\|X\|_F^2, \quad (5)$$

where η is a parameter. We will further discuss the convexity of $f(X)$ in Section 4.

3.3. The FDDL Model

By incorporating Eqs. (4) and (5) into Eq. (3), we have the following FDDL model:

$$J_{(D,X)}=\arg\min_{(D,X)}\left\{\begin{array}{l}\sum_{i=1}^{c}r(A_i,D,X_i)+\lambda_1\|X\|_1+\\ \lambda_2\left(tr(S_W(X)-S_B(X))+\eta\|X\|_F^2\right)\end{array}\right\}. \quad (6)$$

Although the objective function J in Eq. (6) is not jointly convex to (D, X), it is convex with respect to each of D and X when the other is fixed. Therefore, an algorithm of alternatively optimizing D and X can be designed. Detailed optimization procedures are presented next in Section 4.

4. Optimization of FDDL

The FDDL objective function in Eq. (6) can be divided into two sub-problems: updating X by fixing D; and updating D by fixing X. The procedures are iteratively implemented for the desired discriminative dictionary D and the discriminative coefficients X.

First, suppose that D is fixed, and the objective function $J_{(D,X)}$ in Eq. (6) is reduced to a sparse coding problem to compute $X=[X_1, X_2, \ldots, X_c]$. Here we compute X_i class by class. When compute X_i, all X_j, $j\neq i$, are fixed. Thus the objective function in Eq. (6) is further reduced to:

$$J_{(X_i)}=\arg\min_{(X_i)}\{r(A_i,D,X_i)+\lambda_1\|X_i\|_1+\lambda_2 f_i(X_i)\} \quad (7)$$

with

$$f_i(X_i)=\|X_i-M_i\|_F^2-\sum_{k=1}^{c}\|M_k-M\|_F^2+\eta\|X_i\|_F^2 ,$$

where M_k and M are the mean vector matrices (by taking n_k mean vectors $\boldsymbol{m}_k$ or $\boldsymbol{m}$ as its column vectors) of class k and all classes, respectively. It can be proved that if $\eta>1-n_i/n$, $f_i(X_i)$ is strictly convex to X_i (please refer to **Appendix A**), where n_i and n are the number of training samples in the i^{th} class and all classes, respectively. In order to make $f_i(X_i)$ not only convex but also have enough discrimination, we set $\eta=1$. Then we can see that all the terms in Eq. (7), except for $\|X_i\|_1$, are differentiable, and Eq. (7) is strictly convex. The Iterative Projection Method (IPM) in [19] can be employed to solve Eq. (7), as described in **Appendix B**.

When X is fixed, we update D_i class by class. When update D_i, all $D_j, j\neq i$, are fixed. Now the objective function in Eq. (6) is reduced to:

$$J_{(D_i)}=\arg\min_{(D_i)}\left\{\begin{array}{l}\left\|A-D_iX^i-\sum_{j=1,j\neq i}^{c}D_jX^j\right\|_F^2+\\ \|A_i-D_iX_i^i\|_F^2+\sum_{j=1,j\neq i}^{c}\|D_iX_j^i\|_F^2\end{array}\right\}, \quad (8)$$

where X^i is the coding coefficients of A over D_i.

In general, we require that each column of the dictionary D_i, denoted by $\boldsymbol{d}_i^l$, is a unit vector. Eq. (8) is a quadratic programming problem and it can be solved by using the algorithm in [13], which updates D_i atom by atom.

Table 1: Algorithm of Fisher Discrimination Dictionary Learning

Fisher Discrimination Dictionary Learning
1. Initialization *D*. We initialize all the p_i atoms of each D_i as random vectors with unit l_2-norm.
2. Update the sparse coding coefficients *X*. Fix D and solve X_i, i=1,2,…,c, one by one by solving Eq. (7) with the algorithm in **Table 7** in **Appendix B**.
3. Updating dictionary *D*. Fix X and update each D_i, i=1,2,…,c, by solving Eq. (8) with the method presented in [13].
4. Output. Return to **step 2** until the values of $J_{(D,X)}$ in adjacent iterations are close enough, or the maximum number of iterations is reached. Output X and D.

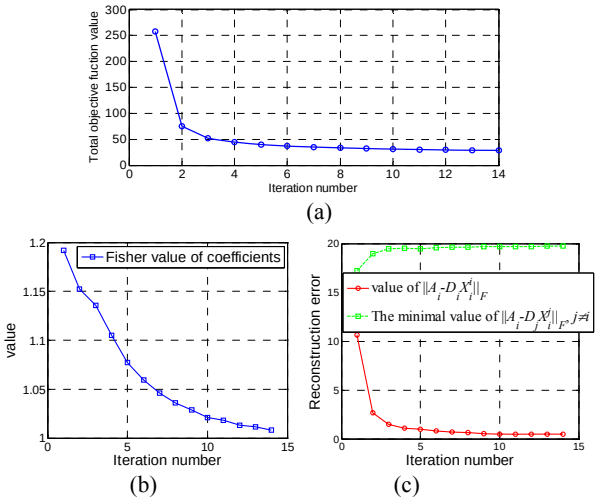

Figure 2: An example of FDDL process on the Extended Yale B face database. (a) The convergence of FDDL. (b) The curve of $tr(S_W(X))/tr(S_B(X))$ versus iteration number. (c) The curves of the reconstruction error of D_i to A_i and the minimal reconstruction error of D_j to A_i, $j\neq i$, versus the iteration number.

The algorithm of FDDL is summarized in Table 1. FDDL converges since the two alternative optimizations in it are both convex. Fig. 2(a) illustrates the convergence of FDDL. Fig. 2(b) shows that the ratio $tr(S_W(X))/tr(S_B(X))$ (essentially the same as $tr(S_W(X))-tr(S_B(X))$ in representing the discrimination ability of X but is invariant to the scale of X) decreases as the iteration number increases, which indicates that X is discriminative after learning the dictionary D. Fig. 2(c) plots the curves of $\|A_i-D_iX_i^i\|_F$ (i=10 here) and the minimal value of $\|A_i-D_jX_i^j\|_F$, j=1,2,…,c, $j\neq i$, showing that D_i could represent A_i well, but D_j, $j\neq i$, has poor representation ability to A_i.

With FDDL, we could use the sparse coding coefficients of each class, i.e., X_i, to compute the mean coefficient vector of that class, denoted as $\boldsymbol{m}_i$, which will then be used for the testing sample classification.

5. The classification scheme

When D is available, a testing sample can be classified via coding it over D. Based on the employed dictionary D, different information can be utilized for the classification task. In [16] and [10], a common dictionary is shared by all classes, and the sparse coding coefficients are used for classification. In SRC [5], the original training samples are used to form a structured dictionary to code the testing sample, and the reconstruction error associated with each class is used for classification. Compared to SRC, in [14] and [11] the testing sample is coded on each sub-dictionary associated with each class, and then the reconstruction error is computed for classification.

Although the methods in [5, 10, 11, 14, 16] lead to good results, they cannot use both the reconstruction error and the coding coefficient for image classification. With the proposed FDDL, however, the learned dictionary D will make both the reconstruction error and the coding coefficient discriminative. Naturally, we can make use of both of them for more accurate classification results. According to the situation of training samples, we propose two classification schemes, the global classifier (GC) and local classifier (LC).

1) GC: When the number of training samples of each class is relatively small, the learned dictionary D_i may not be able to faithfully represent the testing samples of this class, and hence we code the testing sample $\boldsymbol{y}$ over the whole dictionary D. In this case, the sparse coding coefficients could be got by solving

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2 + \gamma \|\boldsymbol{\alpha}\|_1 \right\}, \qquad (9)$$

where γ is a constant. Denote by $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_c]$, where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with sub-dictionary D_i. We define the metric for final classification as

$$e_i = \|\boldsymbol{y} - D_i\hat{\boldsymbol{\alpha}}_i\|_2^2 + w \cdot \|\hat{\boldsymbol{\alpha}} - \boldsymbol{m}_i\|_2^2, \qquad (10)$$

where the first term is the reconstruction error by class i, the second term is the distance between the coefficient vector $\hat{\boldsymbol{\alpha}}$ and the learned mean vector $\boldsymbol{m}_i$ of class i, and w is a preset weight to balance the contribution of the two terms. The classification of $\boldsymbol{y}$ is made by Eq. (2).

2) LC: When the number of training samples of each class is relatively large, the learned dictionary D_i is able to well span the sample space of class i, and thus we could directly code $\boldsymbol{y}$ by D_i to reduce the computational cost and the interference of other dictionaries. Denote by $\boldsymbol{m}_i$ =

$[\boldsymbol{m}_i^1;\ldots;\boldsymbol{m}_i^k;\ldots;\boldsymbol{m}_i^c]$, where $\boldsymbol{m}_i^k$ is the sub-vector associated with sub-dictionary D_k. The coding coefficients associated with D_i are got by solving

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y} - D_i\boldsymbol{\alpha} \right\|_2^2 + \gamma_1 \left\| \boldsymbol{\alpha} \right\|_1 + \gamma_2 \left\| \boldsymbol{\alpha} - \boldsymbol{m}_i^i \right\|_2^2 \right\}, \quad (11)$$

where γ_1 and γ_2 are constants. Here we require that not only D_i should well code $\boldsymbol{y}$ with sparse coefficients, but also the coding vector $\boldsymbol{\alpha}$ should be close to $\boldsymbol{m}_i^i$, the i^{th}-class trained mean vector associated with sub-dictionary D_i. Hence the metric for final classification is defined as

$$e_i = \left\| \boldsymbol{y} - D_i\hat{\boldsymbol{\alpha}} \right\|_2^2 + \gamma_1 \left\| \hat{\boldsymbol{\alpha}} \right\|_1 + \gamma_2 \left\| \hat{\boldsymbol{\alpha}} - \boldsymbol{m}_i^i \right\|_2^2. \quad (12)$$

The final classification rule is the same as Eq. (2).

6. Experimental results

We verify the performance of FDDL on applications such as FR, digit recognition and gender classification. Since the number of training samples is often small in FR, we used GC as the classifier. For digit recognition, each class has many training samples and thus we used LC as the classifier. For gender classification, we tested both the classifiers. The source code of FDDL can be found at http://www4.comp.polyu.edu.hk/~cslzhang/code.htm.

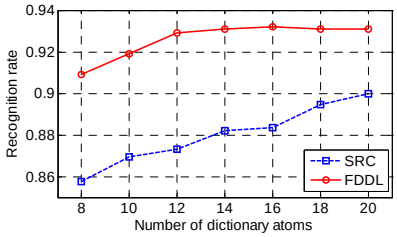

Figure 3: The recognition rates of FDDL and SRC versus the number of dictionary atoms.

6.1. Parameter selection

One important parameter in FDDL is the number of atoms in D_i, denoted by p_i. For FDDL, we usually set all the p_i equal, i=1,2,…c. We use SRC as the baseline method, and analyze the effect of p_i on the performance of FDDL. We take FR on Extended Yale B [21-22] as an example (the experiment setting is given in next subsection). Because SRC uses the original training samples as dictionary, we randomly select p_i training samples as dictionary atoms and run 10 times the experiment to get the average recognition rate. Fig. 3 plots the recognition rates of FDDL and SRC versus different number of dictionary atoms. We can see that in all cases FDDL has about 3% improvement over SRC. Especially, even with the atom number 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, FDDL's recognition rate drops by 2.2%, compared to 4.2% for SRC. This shows that FDDL is able to compute a compact and representative dictionary, which has less computational cost and higher recognition rate than SRC. The time complexity of FDDL classification is comparable to DLSI, and a little higher than DKSVD (shared dictionary with a smaller size).

In all experiments, if no specific instructions, the tuning parameters in FDDL (λ_1 and λ_2 in dictionary learning phase, γ and w in GC or γ_1 and γ_2 in LC) and the parameters of competing methods are evaluated by 5-fold cross validation to avoid over-fitting.

6.2. Face recognition

We apply the proposed algorithm to FR on the Extended Yale B [21-22], AR [23], and Multi-PIE [24] face databases. In order to clearly illustrate the advantage of the proposed method, we compare FDDL with SRC, two latest DL based classification methods (*discriminative KSVD* (DKSVD) [10] and *dictionary learning with structure incoherence* (DLSI) [11]) and two popular classification methods (*nearest neighbor* (NN) and linear *support vector machines* (SVM)). Note that the original DLSI method codes the testing sample by each class. For a fair comparison, we also gave the results (denoted by DLSI*) by coding the testing sample over the whole dictionary and using the reconstruction error for classification. The default number of dictionary atoms in FDDL on each class is set as the number of training samples. The Eigenface [25] with dimension 300 is used in all FR experiments, and the parameters of FDDL chosen by cross-validation are λ_1=0.005, λ_2=0.005, γ=0.001, w=0.05 for Extended Yale B, λ_1=0.005, λ_2=0.05, γ=0.005 and w=0.05 for AR, and λ_1=0.005, λ_2=0.05, γ=0.01 and w= 1 (0.5) for MPIE Test 1 (Test 2).

a) Extended Yale B: The Extended Yale B database consists of 2,414 frontal-face images from 38 individuals (about 64 images per subject) captured under various laboratory-controlled lighting conditions. For each subject, we randomly selected 20 images for training with the remaining images for testing (the experimental setting is more difficult than that in [5]). The images were normalized to 54×48. The results of FDDL, SRC, NN, SVM, DKSVD and DLSI are listed in Table 2. It can be seen that FDDL improves at least 2% over the other methods. DKSVD, which only uses coding coefficients for classification, does not work well. DLSI* has better results than DLSI, showing that coding the query image on the whole dictionary is more reasonable in this case.

b) AR: The AR database consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separated sessions. As in [5], in the experiment we chose a subset consisting of 50 male subjects and 50 female subjects. 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 original face image is 60×43. The comparison between competing methods is shown in Table 3. Again, FDDL has at least 3% improvement over the other methods. DLSI* has the second best performance; however, DLSI gets the second worst results because each class has only 7 training samples in this experiment.

Table 2: The recognition rates of various methods on the Extended Yale B database.

Method	SRC	NN	SVM	DKSVD	DLSI DLSI*	**FDDL**
Recognition rate	0.900	0.617	0.888	0.753	0.850 0.890*	**0.919**

Table 3: The recognition rates of various methods on the AR database.

Method	SRC	NN	SVM	DKSVD	DLSI DLSI*	**FDDL**
Recognition rate	0.888	0.714	0.871	0.854	0.737 0.898*	**0.920**

Table 4: The recognition rates of various methods on the Multi-PIE database.

Method	SRC	NN	SVM	DKSVD	DLSI DLSI*	**FDDL**
Test 1	0.955	0.902	0.916	0.939	0.914 0.941*	**0.967**
Test 2	0.961	0.947	0.922	0.898	0.949 0.959*	**0.980**

Table 5: Error rates of various methods on digit recognition.

Algorithms	Error rate (%)
FDDL	**3.69**
SRSC	6.05
REC-L	6.83
REC-BL	4.38
SDL-G	6.67
SDL-D	**3.54**
DLSI	3.98
KNN	5.2
SVM-Gauss	4.2

c) Multi-PIE: The CMU Multi-PIE face database [24] is a large scale database of 337 subjects including four sessions with simultaneous variations of pose, expression and illumination. Among the 337 subjects, we chose 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 different methods are listed in Table 4. We can see that compared with the previous methods, FDDL has at least 1% (in Test 1) or 2% (in Test 2) improvement with a smaller dictionary. SRC works the second best.

In all the FR experiments, DLSI* advances DLSI, and DKSVD is worse than FDDL, SRC and DLSI*, which may imply that the reconstruction error associated with each class is more powerful than the coding coefficients in face classification.

6.3. Digit recognition

We then perform handwritten digit recognition on the widely used USPS database [26] with 7,291 training and 2,007 testing images. We compare the proposed FDDL with state-of-the-art methods reported in [11], [9] and [8]. These methods include the best reconstructive DL method with linear and bilinear classifier models (denoted by REC-L and REC-BL) [9], the best supervised DL method with generative training and discriminative training (denoted by SDL-G and SDL-D) [9], the best result of sparse representation for signal classification (denoted by SRSC) [8] and the best result of DLSI [11]. In addition, some results of problem-specific methods (i.e., the standard Euclidean k_NN and SVM with a Gaussian kernel) reported in [11] are also listed. Here the original image (16×16) is directly used as the feature and the dictionary of each class has 90 atoms in FDDL with $\lambda_1 = \gamma_1 = 0.1$, $\lambda_2 = 0.001$, and $\gamma_2 = 0.005$.

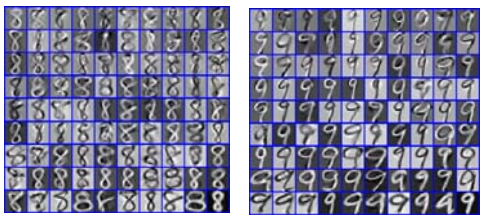

Figure 4: The learned bases of digits 8 and 9 by FDDL.

Fig. 4 illustrates the learned bases of digits 8 and 9. Table 5 lists the results of FDDL and its competing methods. We see that FDDL outperforms all the competing methods except for SDL-D (FDDL and SDL-D have very close results). It should be noted that SDL-D uses more information in DL and classification processes, including a learnt classifier of coding coefficients, the sparsity of coefficients, and the reconstruction error. In addition, the optimization of SDL-D method is much more complex than that of FDDL.

[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}.

6.4. Gender classification

In this experiment we chose a non-occluded subset (14 images per subject) of AR consisting of 50 male subjects and 50 female subjects. Images of the first 25 males and 25 females were used for training, and images of the remaining 25 males and 25 females for testing. We used PCA to reduce the dimension of each image to 300. In addition, we present the result of DLSI# (coding on the whole dictionary and classifying like SRC). Here p_i is set as 250 for FDDL and RBF kernel is adopted in SVM.

Table 6 lists the recognition results of FDDL and the competing methods. It can be seen that FDDL with LC gets the best result when coding the testing image by the dictionary of each class (1.4% improvement over the second best one, DLSI); while FDDL with GC gets the best result when coding the test sample over the whole dictionary (1.1% higher than the second best one, SRC). Meanwhile, we can see that DLSI and FDDL with LC have better performance than DLSI# and FDDL with GC, respectively. 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 for the testing sample.

Table 6: The results of different methods on gender classification using the AR database.

SRC	DK-SVD	DLSI DLSI#	**FDDL with LC** **FDDL with GC**	SVM	NN
0.930	0.861	0.940 0.900	**0.954** **0.941**	0.924	0.907

7. Conclusion and discussion

In this paper, we proposed a Fisher Discrimination Dictionary Learning (FDDL) approach to sparse representation based image classification. The FDDL aims to learn a structured dictionary whose sub-dictionaries have specific class labels. The discrimination ability of FDDL is two-folds. First, each sub-dictionary of the learned whole dictionary has good representation power to the samples from the corresponding class, but has 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 error and sparse coding coefficients to classify the input query image. The experimental results on face recognition (FR), digit recognition and gender classification clearly demonstrated the superiority of FDDL to many state-of-the-art dictionary learning based methods. In future, we will apply FDDL to other classification tasks such as object recognition.

Very recently, Zhang *et al.* [32] indicated that it is the collaboratively representation strategy in SRC, but not the l_1-norm sparsity constraint, that truly helps FR. Based on our experimental experience, removing $\|X\|_1$ and preserving only the term $\|X\|_F^2$ in the FDDL model in Eq. (6), and replacing $|\boldsymbol{\alpha}|_1$ by $\|\boldsymbol{\alpha}\|_2^2$ in Eqs. (9) and (11) can indeed lead to similar FR results but with much less computational cost. However, it is not clear yet if this is true to other pattern classification tasks. It needs more investigation about the role of l_1-norm sparsity in dictionary learning.

References

[1] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. IEEE TIP, 15(12):3736–3745, 2006.

[2] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. IEEE TIP, 17(1):53–69, 2008.

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

[4] E. Candes. Compressive sampling. Int. Congress of Mathematics, 3:1433–1452, 2006.

[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. IEEE TPAMI, 31(2):210–227, 2009.

[6] A. Wagner, J. Wright, A. Ganesh, Z.H. Zhou, and Y. Ma. Towards a Practical Face Recognition System: Robust Registration and Illumination by Sparse Representation. In CVPR, 2009.

[7] M. Yang and L. Zhang. Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In ECCV, 2010.

[8] K. Huang and S. Aviyente. Sparse representation for signal classification. In NIPS, 2006.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In NIPS, 2009.

[10] Q. Zhang and B.X. Li. Discriminative K-SVD for dictionary learning in face recognition. In CVPR, 2010.

[11] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In CVPR, 2010.

[12] J.C. Yang, K. Yu, and T. Huang. Supervised Translation-Invariant Sparse coding. In CVPR, 2010.

[13] M. Yang, L. Zhang, J. Yang and D. Zhang. Metaface learning for sparse representation based face recognition. In ICIP, 2010.

[14] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zissserman. Learning discriminative dictionaries for local image analysis. In CVPR, 2008.

[15] F. Rodriguez and G. Sapiro. Sparse representation for image classification: Learning discriminative and reconstructive non-parametric dictionaries. IMA Preprint 2213, 2007.

[16] D. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In CVPR, 2008.

[17] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE TSP, 54(11):4311–4322, 2006.

[18] R. Duda, P. Hart, and D. Stork. Pattern classification (2nd ed.), Wiley-Interscience, 2000.

[19] L. Rosasco, A. Verri, M. Santoro, S. Mosci, and S. Villa. Iterative Projection Methods for Structured Sparsity Regularization. MIT Technical Reports, MIT-CSAIL-TR-2009-050, CBCL-282, 2009.
[20] J. Mairal, M. Leordeanu, F. Bach, M. Hebert and J. Ponce. Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation. In ECCV, 2008.
[21] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. IEEE TPAMI, 27(5): 684–698, 2005.
[22] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE TPAMI, 23(6):643–660, 2001.
[23] A. Martinez and R. benavente. The AR face database. CVC Tech. Report No. 24, 1998.
[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. Image and Vision Computing. 28:807–813, 2010.
[25] M. Turk and A. Pentland. Eigenfaces for recognition. J. Cognitive Neuroscience, 3(1):71–86, 1991.
[26] USPS Handwritten Digit Database. *http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html.*
[27] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004.
[28] J.A. Tropp and S.J. Wright. Computational methods for sparse solution of linear inverse problems. Proc. of IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation, 98(6):948–958, 2010.
[29] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM. J. Imaging Science, 2(1):183-202, 2009.
[30] B.A. Olshausen, D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381:607-609, 1996.
[31] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In CVPR, 2011.
[32] L. Zhang, M. Yang, and X.C. Feng. Sparse representation or collaborative representation: which helps face recognition? In ICCV 2011.

Appendix A: The convexity of $f_i(X)$

Let $E_i^j = [1]_{n_i \times n_j}$ be a matrix of size $n_i \times n_j$ with all entries being 1, and let $N_i = I_{n_i \times n_i} - E_i^i / n_i$, $P_i = E_i^i / n_i - E_i^i / n$, $C_i^j = E_i^j / n$.

From $f_i(X_i) = \|X_i - M_i\|_F^2 - \sum_{k=1}^{c} \|M_k - M\|_F^2 + \eta \|X_i\|_F^2$, we can derive that

$$f_i(X_i) = \|X_i N_i\|_F^2 - \|X_i P_i - G\|_F^2 - \sum_{\substack{k=1, \\ k \neq i}}^{c} \|Z - X_i C_i^k\|_F^2 + \eta \|X_i\|_F^2, \quad (13)$$

where $G = \sum_{k=1,k\neq i}^{c} X_k C_k^i$, $Z = X_k E_k^k / n_k - \sum_{j=1,j\neq i}^{c} X_j C_j^k$.

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

$$f_i(\boldsymbol{\chi}_i) = \left\|\mathrm{diag}(N_i^T)\boldsymbol{\chi}_i\right\|_2^2 - \left\|\mathrm{diag}(P_i^T)\boldsymbol{\chi}_i - \mathrm{diag}(G^T)\right\|_2^2 - \sum_{k=1,k\neq i}^{c} \left\|\mathrm{diag}\left((C_i^k)^T\right)\boldsymbol{\chi}_i - \mathrm{diag}(Z^T)\right\|_2^2 + \eta\|\boldsymbol{\chi}_i\|_2^2,$$

where diag(T) is to construct a block diagonal matrix with each block on the diagonal being matrix 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 [27]. $\nabla^2 f_i(\boldsymbol{\chi}_i)$ will be positive definite if the following matrix S is positive definite:

$$S = N_i N_i^T - \left(P_i P_i^T + \sum_{k=1,k\neq i}^{c} C_i^k \left(C_i^k\right)^T\right) + \eta I \ .$$

After some derivations, we have

$$S = (1+\eta) I - E_i^i \left(2/n_i - 2/n + \sum_{k=1}^{c} n_k / n^2\right).$$

In order to make S positive define, each eigenvalue of S should be greater than 0. Because the maximal eigenvalue of E_i^i is n_i, we should ensure

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

For $n=n_1+n_2+\ldots+n_c$, we have $\eta > 1-n_i/n$, which could guarantee that $f_i(X_i)$ is convex to X_i.

Appendix B: The coding algorithm of FDDL

We rewrite Eq. (7) as

$$J_{(X_i)} = \arg\min_{(X_i)} \left\{Q(X_i) + 2\tau \|X_i\|_1\right\} \quad (14)$$

where $Q(X_i) = r(A_i, D, X_i) + \lambda_2 f_i(X_i)$, $\tau=\lambda_1/2$, and $f_i(X_i)$ and $r(A_i,D,X_i)$ are defined in Eqs. (4) and (7), respectively.

Define $\tilde{X}_i = \left[\boldsymbol{x}_{i,1}^T, \boldsymbol{x}_{i,2}^T, \cdots, \boldsymbol{x}_{i,n_i}^T\right]^T$, where $\boldsymbol{x}_{i,k}$ is the k^{th} column vector of matrix X_i. Because $Q(X_i)$ is strictly convex to X_i, the sparse coding problem in Eq. (14) could be solved by the Iterative Projective Method [19]. Table 7 describes the algorithm of minimizing Eq. (14), whose speed could be improved by approaches like FISTA [29].

Table 7: The coding algorithm of FDDL

Coding algorithm of FDDL
1. **Input**: σ, τ >0.
2. **Initialization**: $\tilde{X}_i^{(1)} = \mathbf{0}$ and h=1.
3. **While** convergence and the maximal iteration number are not reached **do** $h = h+1$ $\tilde{X}_i^{(h)} = \boldsymbol{S}_{\tau/\sigma}\left(\tilde{X}_i^{(h-1)} - \frac{1}{2\sigma}\nabla Q\left(\tilde{X}_i^{(h-1)}\right)\right)$ where $\nabla Q\left(\tilde{X}_i^{(h-1)}\right)$ is the derivative of $Q(X_i)$ w.r.t. $\tilde{X}_i^{(h-1)}$, and $\boldsymbol{S}_{\tau/\sigma}$ is a soft thresholding function defined in [19].
4. **Return** $\tilde{X}_i = \tilde{X}_i^{(h)}$.