A New Approach for Learning Discriminative Dictionary for Pattern Classification

THUY THI NGUYEN[1], BINH THANH HUYNH[2] AND SANG VIET DINH[2]
[1]*Faculty of Information Technology*
Vietnam National University of Agriculture
Trau Quy town, Gialam, Hanoi, Vietnam
E-mail: myngthuy@gmail.com
[2]*School of Information and Communication Technology*
Hanoi University of Science and Technology
No 1, Dai Co Viet Street, Hanoi, Vietnam
E-mail: {binhht; sangdv}@soict.hust.edu.vn

Dictionary learning (DL) for sparse coding based classification has been widely researched in pattern recognition in recent years. Most of the DL approaches focused on the reconstruction performance and the discriminative capability of the learned dictionary. This paper proposes a new method for learning discriminative dictionary for sparse representation based classification, called Incoherent Fisher Discrimination Dictionary Learning (IFDDL). IFDDL combines the Fisher Discrimination Dictionary Learning (FDDL) method, which learns a structured dictionary where the class labels and the discrimination criterion are exploited, and the Incoherent Dictionary Learning (IDL) method, which learns a dictionary where the mutual incoherence between pairs of atoms is exploited. In the combination, instead of considering the incoherence between atoms in a single shared dictionary as in IDL, we propose to incorporate the incoherence between pairs of atoms within each sub-dictionary, which represent a specific object class. This aims to increase discrimination capacity of between basic atoms in sub-dictionaries. The combination allows one to exploit the advantages of both methods and the discrimination capacity of the entire dictionary. Extensive experiments have been conducted on benchmark image data sets for Face recognition (ORL database, Extended Yale B database, AR database) and Digit recognition (the USPS database). The experimental results show that our proposed method outperforms most of state-of-the-art methods for sparse coding and DL based classification, meanwhile maintaining similar complexity.

Keywords: dictionary learning, sparse coding, fisher criterion, pattern recognition, object classification

1. INTRODUCTION

Sparse representation (or sparse coding) has been widely used in many problems of image processing and computer vision [1, 2], audio processing [3, 4], as well as classification [5-9] and archived very impressive results. In this model, an input signal is decomposed by a sparse linear combination of a few atoms from an over-complete dictionary. In general, the goal of sparse representation is to represent input signals by a linear combination of atoms (or words). This is done by minimizing the reconstruction error under a sparsity constraint:

$$\min_{D,X} \left(\| A - DX \|_F^2 + \lambda \| X \|_1 \right) \quad (1)$$

Received February 15, 2015; revised June 18, 2015; accepted July 9, 2015.
Communicated by Hsin-Min Wang.

where $A = [a_1, a_2, ..., a_n] \in \mathbb{R}^{m \times n}$ is a set of n training samples $a_i \in \mathbb{R}^m, i = 1, 2, ..., n$; $D = [d_1, d_2, ..., d_p] \in \mathbb{R}^{m \times p}$ is the over-complete dictionary to be learned, containing p atoms ($p > m$ and $p \ll n$); $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{p \times n}$ is the coefficient matrix consisting of n sparse coding vectors $x_i \in \mathbb{R}^p, i = 1, 2, ..., n$; and λ is a parameter to balance the reconstruction error and the sparsity level.

With the over-complete property, the learned dictionary can discover interesting features in the data and often provide good results in many tasks. Besides that, the sparsity property gives an efficient way to store information of the signals.

Because the sparsity constraint in Eq. (1) is not jointly convex with respect to both D and X , but is convex with respect to each of the two variables when other one is fixed, the problem is usually solved by iteratively optimizing two sub-problems: sparse coding and dictionary learning. In the sparse coding sub-problem, the coefficient matrix X is estimated while keeping the dictionary D fixed by using some algorithm like Matching Pursuit (MP) [10] or Orthogonal Matching Pursuit (OMP) [11]. In the dictionary learning sub-problem, dictionary D is updated while the coefficient matrix X is fixed. Some well-known methods to update D are MOD [12] and KSVD [13]. In MOD method, dictionary is updated by using least square solution of the problem in Eq. (1). In another approach, K-SVD updates each atom one by one by using rank-one matrix approximation of residual matrix. Both MOD and K-SVD give the similar result, K-SVD often gives a faster convergence speed than MOD.

Although learned dictionaries can give good approximations in signal presentations, they have some drawback in classification tasks. To make the dictionary more discriminative, many supervised dictionary learning methods have been proposed based on the basic sparse model. Zhang and Li [14] added classification error to the objective function of dictionary learning. Another approach is adding discriminative sparse-code error term [5] to make sure that each sub-dictionary of each class should be good at presenting the samples from this class and not good for other classes. Yang et al. [15] proposed a stronger discriminative dictionary learning method using Fisher criterion called Fisher Discrimination Dictionary Learning (FDDL). With the Fisher discrimination criterion, the dictionary obtained by this algorithm forces the sparse codes of samples from the same class to be similar while the sparse codes of samples from different classes to be far enough. Experiments have shown that FDDL can give the state-of-the-art results in dictionary learning for classification. Another approach in learning dictionary was proposed in [6] called Incoherent Dictionary Learning (IDL). This approach forces an incoherent promoting term to make the pairs of atoms of the dictionary as orthonormal to each other as possible. Without using the class labels, IDL still can make the learned dictionary powerful in classification tasks.

Although FDDL gives good discrimination between sub-dictionaries, it does not take into account the discrimination between atoms within each sub-dictionary or the incoherence between them. In this paper, we propose an improved version of FDDL by forcing the incoherence of atoms within each sub-dictionary. This is done by modifying the objective function of the FDDL model, where the incoherent promoting term is added. By doing so, the novel IFDDL model imposes an incoherent constraint on each sub-dictionary to minimize the correlation between the atoms belong to it.

The paper is organized as follows: In section 2 we briefly review related work on dictionary learning. In section 3, we present the proposed algorithm for dictionary learn-

ing. Our experiments and evaluation are in section 4. The conclusion is in section 5 with a discussion and the future work.

2. RELATED WORK

The goal of DL is to create from the training set a group of atoms that can well represent the given samples. Over the last years, a large amount of DL methods has been proposed, especially in the field of image processing [13, 16] and object recognition [6, 8, 9, 14, 17-20]. One of the most efficient methods for DL is the KSVD algorithm [13], which is a generalization of the K-means algorithm. However, when DL is used for classification tasks, the discrimination capability of the over-complete dictionary obtained by KSVD is not guaranteed. Therefore, it has drawback for the problem of pattern classification or object recognition.

Based on the spirit of KSVD, in order to enhance the discrimination capability of the learned dictionary, a discriminative reconstruction constraint was added to the objective function [17]. However, the objective function of this method is not convex and it does not exploit the discrimination capability of the coefficient matrix. Pham and Venkatesh [18] proposed a joint learning and dictionary construction method by formulating an objective function that combines classification error with representation errors of both labeled and unlabeled data, and applied their method to object categorization and face recognition (FR). Based on [18], Zhang and Li [14] proposed the so-called discriminative KSVD (DKSVD) algorithm using only labeled data in dictionary learning procedure. The authors in [5] proposed a new discriminative DL method by training a classifier of the coding coefficients. All the above mentioned methods learn a common dictionary for all classes and a classifier of coefficients for classification. However, the single shared dictionary does not contain the information about the correspondence between the atoms and the class labels that would allow increasing classification performance.

There are a number of approaches where the main idea is to learn a sub-dictionary for each class using the correspondence between the atoms and the class labels. One of the most efficient methods was proposed in [5] using label consistent information. Compared to the previous methods using the same idea, this method has significantly improved FR results. Ramirez et al. [8] used an incoherence constraint to make the sub-dictionaries as independent as possible. However, these methods did not impose any discrimination constraint on the coefficient matrix X. In order to make the coding coefficients discriminative Yang et al. [15] proposed a new method using Fisher discriminative criterion called Fisher Discriminative Dictionary Learning (FDDL).

FDDL used the training data set of each class to learn a sub-dictionary for that class. After all sub-dictionaries for all class are learned, a large common dictionary is created by combining all sub-dictionaries. Besides, FDDL added some constraints to the objective function to guarantee the sub-dictionary of class i only well represents the data set of the class i. In addition, FDDL imposed the Fisher discrimination criterion on the coefficient matrix X to minimize the within-class scatter of X meanwhile maximizing the between-class scatter. This method has shown impressive result on some data sets. Despite of that, each sub-dictionary built by this method is independent, and the incoherent information of atoms in each sub-dictionary has not yet been exploited.

The IDL method in [6] incorporated the mutual incoherence between pairs of atoms into the learning process to increase the discrimination capacity of the learned dictionary. Besides that, IDL employed supervised learning approach, i.e. using information about class labels, to improve the accuracy of classification tasks. Although the discrimination capability of dictionary in IDL is lower than FDDL in the reported experiments, the mutual incoherence between the atoms is an important factor to be exploited for learning a good dictionary.

In another approach, there have been attempts to learn a hybrid dictionary by combining class-specific atoms and common shared dictionary atoms (also called particularity and commonality) such as [21, 22], or learning a latent dictionary [23, 24]. In [21] the discrimination of the dictionaries is enhanced by using Fisher discrimination criterion in a joint learning framework. The incoherence of atoms within the class-specific dictionaries has not yet been investigated. In [22], an incoherence term has been added to the objective function. However, in this work the Fisher discrimination criterion has not been incorporated.

In this work, we will investigate the advantages of the FDDL method for DL. We then explore the correlation between atoms in the sub-dictionary by modifying the objective function of the FDDL [15] using the IDL constraint. This allows us to add the incoherent information of atoms for a better representation but still remain the advantages of both FDDL and IDL.

3. OUR PROPOSED APPROACH

3.1 Preliminary

In the following we will briefly revise techniques used in learning Fisher Discrimination Dictionary Learning (FDDL) and Incoherent Dictionary Learning (IDL) that we will base on to build our algorithm.

3.1.1 Fisher Discrimination Dictionary Learning

The main idea of FDDL is based on the classical sparse representation problem in Eq. (1). However, instead of minimizing only reconstruction error for whole dictionary under a sparsity constraint as in Eq. (1), FDDL uses additional reconstruction errors corresponding to different class-specified sub-dictionaries. Furthermore, FDDL imposes a constraint on the coefficient matrix to make the learned dictionary more discriminative.

FDDL learns a structured dictionary $D = [D_1, D_2, ..., D_K]$, where $D_i \in \mathbb{R}^{m \times p_i}$ is the sub-dictionary corresponding to class i; p_i is the number of atoms in sub-dictionary D_i; and K is the number of classes.

Suppose that the data matrix A is decomposed to be a set of sub-matrices $A = [A_1, A_2, ..., A_K]$, where A_i is the subset of training samples associated with class i. Let $X_i, i = 1, 2, ..., K$ be the representation matrix of A_i over D_i, then X can be expressed as $X = [X_1, X_2, ..., X_K]$. The objective function of the FDDL model is:

$$J(D, X) = r(A, D, X) + \lambda_1 \| X \|_1 + \lambda_2 f(X), \quad (2)$$

Where $r(A, D, X)$ is the residual component that characterizes the representation capability of the learned dictionary D to the training set A; $\| X \|_1$ is the l_1-norm sparsity regularization of the coefficient matrix; $f(X)$ is a Fisher discrimination-based constraint imposed on the coefficient matrix X to enhance the discriminative capability of coding vectors; and λ_1, λ_2 are scalar hyper parameters which are often tuned using cross-validation.

- The residual component $r(A, D, X)$

Suppose that X_i, the representation matrix of A_i over the entire dictionary D, is decomposed as $X_i = [X_i^1, ..., X_i^j, ..., X_i^K]$, where X_i^j is the representation matrix of A_i over the sub-dictionary D_j.

The residual component $r(A, D, X)$ is the sum of class-specified residual terms $r(A_i, D, X_i)$:

$$r(A, D, X) = \sum_{i=1}^{K} r(A_i, D, X_i), \tag{3}$$

where

$$r(A_i, D, X_i) = \| A_i - DX_i \|_F^2 + \| A_i - DX_i^i \|_F^2 + \sum_{\substack{j=1 \\ j \neq i}}^{K} \| D_j X_i^j \|_F^2 . \tag{4}$$

The first term $\| A_i - DX_i \|_F^2$ in Eq. (4) makes sure that the dictionary D can represent well the subset of samples A_i, i.e. $A_i \approx DX_i$. Since A_i should be well represented by the corresponding sub-dictionary D_i, i.e. $A_i \approx D_i X_i^i$, but not by other ones $D_j, j \neq i$, the second term $\| A_i - D_i X_i^i \|_F^2$ in Eq. (4) and the extra representation $\| D_j X_i^j \|_F^2, j \neq i$ should be small.

- The Fisher discrimination-based constraint $f(X)$

To further enhance the discrimination capacity of the dictionary D, one can enforce the coefficient matrix X to be discriminative. In FDDL, the Fisher discrimination-based constraint $f(X)$ encourages the coding vectors of each subset A_i to be as similar as possible, while the coding vectors associated with different classes to be as far from each other as possible. Let $S_W(X)$ be the within-class scatter of X and $S_B(X)$ be the between-class scatter of X:

$$S_W(X) = \sum_{i=1}^{K} \sum_{x_k \in X_i} (x_k - m_i)(x_k - m_i)^T , \tag{5}$$

$$S_B(X) = \sum_{i=1}^{K} n_i (m_i - m)(m_i - m)^T , \tag{6}$$

where m_i and m are the mean vector of X_i and X respectively, n_i is the number of samples in A_i.

The constraint $f(X)$ is defined as:

$$f(X) = tr(S_W(X)) - tr(S_B(X)) + \eta \| X \|_F^2 , \tag{7}$$

Where η is a scalar parameter; and $\| X \|_F^2$ is an elastic regularization which is added to ensure the convexity of $f(X)$.

3.1.2 Incoherent Dictionary Learning

In the incoherent dictionary learning (IDL) [6] an incoherent promoting term is introduced to make the atoms of the learned dictionary as independent as possible. Hence, it contributes to the increasing of the discrimination capacity of the learned dictionary. The incoherent promoting term is defined as a correlation measure between the atoms of D:

$$cor(D) = \| D^T D - I \|_F^2, \tag{8}$$

where $I \in \mathbb{R}^{p \times p}$ is an identity matrix.

The dictionary D is said to be most incoherent if the correlation measure is zero, i.e. all the atoms of D are orthonormal to each other. Minimizing the incoherent term guarantees that the dictionary D can efficiently represent the input samples and achieve higher accuracies for classification tasks.

3.2 Incoherent Fisher discrimination dictionary learning

Despite the fact that FDDL outperforms many state-of-the-art methods in various image classification tasks, it has drawbacks. One of the drawbacks is that it does not consider the correlation between the atoms in the dictionary. Hence, we propose a novel method, called Incoherent Fisher Discrimination Dictionary Learning (IFDDL), to improve the FDDL model in [15] by incorporating the incoherent promoting terms [6].

In IFDDL we impose an incoherent constraint on each sub-dictionary D_i to minimize the correlation between the atoms of D_i. Obviously, one can add a supplementary incoherence constraint, like the one described in [8], to increase the independence of the sub-dictionaries associated with different classes. However, this addition may significantly increase computational complexity of the IFDDL method. The goal of our IFDDL method is to minimize the following objective function with respect to D, X:

$$J(D, X) = r(A, D, X) + \lambda_1 \| X \|_1 + \lambda_2 f(X) + \mu \sum_{i=1}^{K} \| D_i^T D_i - I_i \|_F^2, \tag{9}$$

Where $r(A, D, X)$ and $f(X)$ are defined using the Eq. (3) and Eq. (7); and $I_i \in \mathbb{R}^{p_i \times p_i}$ is the identity matrix corresponding to the sub-dictionary D_i.

Obviously, the objective function $J(D, X)$ can be concretely rewritten as:

$$\begin{aligned} J(D, X) = \sum_{i=1}^{K} \Big(\| A_i - DX_i \|_F^2 + \| A_i - D_i X_i^i \|_F^2 + \sum_{\substack{j=1 \\ j \neq i}}^{K} \| D_i X_j^i \|_F^2 + \mu \| D_i^T D_i - I_i \|_F^2 \Big) + \\ + \lambda_1 \| X \|_1 + \lambda_2 (tr(S_W(X) - S_B(X)) + \eta \| X \|_F^2), \end{aligned} \tag{10}$$

Note that:

$$\sum_{i=1}^{K} \| A_i - DX_i \|_F^2 = \| A - DX \|_F^2 = \| A - D_i X^i - \sum_{\substack{j=1 \\ j \neq i}}^{K} D_j X^j \|_F^2,$$

where X^i is the representation matrix of A over the sub-dictionary D_i.

Consequently, we have:

$$J(D,X) = \| A - D_i X^i - \sum_{\substack{j=1 \\ j\neq i}}^{K} D_j X^j \|_F^2 + \sum_{i=1}^{K} \left(\| A_i - D_i X_i^i \|_F^2 + \sum_{\substack{j=1 \\ j\neq i}}^{K} \| D_i X_j^i \|_F^2 + \mu \| D_i^T D_i - I_i \|_F^2 \right) + + \lambda_1 \| X \|_1 + \lambda_2 \left(tr(S_W(X) - S_B(X)) + \eta \| X \|_F^2 \right). \quad (11)$$

The IFDDL objective function in Eq. (9) can be divided into two sub-problems by learning dictionary and coefficient matrix alternatively: updating X by fixing D, and updating D by fixing X.

In the spare coding sub-problem, we keep D fixed, and update X using the method described in [15].

In the dictionary learning sub-problem, we update D_i class by class while keeping X and other sub-dictionaries $D_j, j \neq i$ fixed. In this case the objective function in Eq. (9) is reduced to:

$$\hat{D}_i = \arg\min_{D_i} \left(\begin{array}{l} \| A - D_i X^i - \sum_{\substack{j=1 \\ j\neq i}}^{K} D_j X^j \|_F^2 + \| A_i - D_i X_i^i \|_F^2 + \\ + \sum_{\substack{j=1 \\ j\neq i}}^{K} \| D_i X_j^i \|_F^2 + \mu \| D_i^T D_i - I_i \|_F^2 \end{array} \right). \quad (12)$$

Let,

$$E_i = A - \sum_{\substack{j=1 \\ j\neq i}}^{K} D_j X^j; L_i = [E_i \; A_i \; \mathbf{0}...\mathbf{0}\mathbf{0}...\mathbf{0}], \quad Z_i = [X^i \; X_i^i \; X_1^i ... X_{i-1}^i X_{i+1}^i ... X_K^i],$$

where each $\mathbf{0}$ is a zero matrix with appropriate size based on the size of $X_j^i, j \neq i$. Then optimization problem in Eq. (12) can be reformulated as:

$$\hat{D}_i = \arg\min_{D_i} \left(\| L_i - D_i Z_i \|_F^2 + \mu \| D_i^T D_i - I_i \|_F^2 \right), \quad (13)$$

Now we will update each atom d_k in D_i one by one, while fixing other atoms $d_l, l \neq k$ in D_i.

Let $Y_k = L_i - \sum_{l\neq k} d_l z_l$, then Eq. (13) becomes:

$$\hat{d}_k = \arg\min_{d_k} \left(\| Y_k - d_k z_k \|_F^2 + \mu \sum_{l\neq k} \left((d_k^T d_l)^2 + (d_l^T d_k)^2 \right) \right), \quad (14)$$

Let

$$F = \| Y_k - d_k z_k \|_F^2 + \mu \sum_{l\neq k} \left((d_k^T d_l)^2 + (d_l^T d_k)^2 \right),$$

then we can take the partial derivative of F with respect to d_k:

$$\nabla_{d_k} F = -2 Y_k z_k^T + 2 d_k z_k z_k^T + 4\mu \sum_{l\neq k} d_l d_l^T d_k,$$

or

$$\nabla_{d_k} F = -2Y_k z_k^T + 2d_k z_k z_k^T + 4\mu D_{ik}^* D_{ik}^{*T} d_k ,$$

where $D_{ik}^* = D_i \setminus d_k , D_{ik}^{*T} = (D_{ik}^*)^T$.

By setting the partial derivative $\nabla_{d_k} F$ to zero, we obtain the update rule for the atom d_k as follows:

$$d_k = (z_k z_k^T I + 2\mu D_{ik}^* D_{ik}^{*T})^{-1} Y_k z_k^T , \tag{15}$$

where $I \in \mathbb{R}^{m \times m}$ is the identity matrix.

Normalize the atom d_k to have unit l_2-norm:

$$d_k = d_k / \| d_k \|_2 . \tag{16}$$

Algorithm 1: Incoherent Fisher Discrimination Dictionary Learning

Input: training sample dataset A and parameters $\mu, \lambda_1, \lambda_2, \eta$.
Output: learned dictionary D and coefficient matrix X.
1: **Initialize dictionary**
In general, we initialize all atoms of D as random vectors with unit l_2-norm. In some specific datasets, some methods such as PCA can be applied.
2: **Update coefficient matrix**
Fix D and update X with the method described in [15].
3: **Update dictionary**
Fix X and update each $D_i, i = 1, 2, ..., K$ one by one:
4: - Update each atom $d_k, k = 1, 2, ..., p_i$ in D_i one by one using Eq. (15) while fixing other atoms $d_l, l \neq k$ in D_i.
5: - Normalize the atom d_k to have unit l_2-norm using Eq. (16).
6: **Repeat**
Return to step 2 until the values of the objective function $J(D, X)$ in Eq. (9) in successive iterations are close enough or the maximum number of iterations is reached.
7: **Output**
Return D and X.

Note that, in general, parameters of a parametric model are natural ones, which reflect the essential properties of the model. However, in many cases, one needs to determine some extra-parameters, for example, to control the configuration of the parametric model. These extra parameters are called hyper parameters. The parameters $\mu, \lambda_1, \lambda_2, \eta$ are added to weight different constraints in the objective function $J(D, X)$, i.e. they are used to regularize $J(D, X)$. In other words, $\mu, \lambda_1, \lambda_2, \eta$ are hyper parameters in our model. As a rule, hyper parameters cannot be directly estimated from a training set. In model selection problem hyper parameters are usually tuned by a k-fold cross validation scheme, typically k =5 or k =10 are usually used. The whole algorithm of the proposed IFDDL method is summarized as in **Algorithm 1**.

3.3 The classification scheme

After the dictionary D is learned, the next step is how to classify a new sample. In our framework, we employ the classification algorithm described in [15]. There are two classification schemes to be explored: the global classifier (GC) and the local classifier (LC). The different between the two schemes is how to compute the sparse codes: in GC scheme, the whole dictionary D is used, while in LC model, the sub-dictionary is used. Because of the property of the used data sets in the experiments, we use the GC scheme.

In GC scheme, the whole dictionary D is used to determine the sparse coding vector of an input sample a as follows:

$$\hat{x} = \arg\min_{x} \left\{ \| a - Dx \|_2^2 + \gamma \| x \|_1 \right\}, \tag{17}$$

Where γ is a scalar constant; $\hat{x} = [x_1; \hat{x}_2; ...; x_p]$ is the coding vector, $\hat{x}_i$ is the coding vector associated with sub-dictionary D_i. We also use residual errors to determine which class one sample should belong to. The residual error of a sparse code vector $\hat{x}_i$ with respect to class i is calculated as:

$$e_i = \| a - D_i \hat{x}_i \|_2^2 + w \| x_i - m_i \|_2^2, \tag{18}$$

where the first term is the reconstruction error by class i, the second term is the distance between the coding vector $\hat{x}_i$ and the learned mean vector m_i of class i, and w is a scalar parameter to scale the contribution of the two terms. The classification of a is made by the following equation:

$$identity(a) = \arg\min_{i} \{ e_i \}. \tag{19}$$

3.4 Computational complexity

Our proposed IFDDL algorithm consists of two main stages: sparse coding and dictionary updating.

Updating the coefficient vector for each sample is a l_1-regularized optimization problem which takes a time approximately of $O(m^2 p^{\varepsilon})$, where $\varepsilon \geq 1.2$ is a constant [25]. Since there are totally n training samples, the computational complexity for sparse coding stage is $O(nm^2 p^{\varepsilon})$.

Note that the sizes of matrices Y_k, z_k, D_{ik}^* in the Eq. (15) are $m \times (2n), 1 \times (2n)$ and $m \times (p_i - 1)$, respectively. For the dictionary updating stage, the time for computing the expression $z_k z_k^T I + 2\mu D_{ik}^* D_{ik}^{*T}$ is approximately $O(2n + m^2(p_i - 1)) \approx O(2n + m^2 p_i))$. The time for computing the expression $Y_k z_k^T$ is approximately $O(2mn)$. In fact, Eq. (15) is a solution of a system of linear equations. While implementing the algorithm in MatLab, instead of directly computing the inverse matrix in Eq. (15), we should use the left division operator, which is based on the Gaussian elimination method, to minimize round-off error. The left division operation typically requires time in an order of $O(m^3)$, because $z_k z_k^T I + 2\mu D_{ik}^* D_{ik}^{*T}$ is a matrix of size $m \times m$.

Hence, the computational complexity for updating each atom in the sub-dictionary D_i is approximately $O(2n + m^2 p_i + 2mn + m^3) \approx O(m^2 p_i + 2mn + m^3)$.

The total computational complexity of dictionary updating stage is approximately

$$\sum_{i=1}^{K} p_i O(m^2 p_i + 2mn + m^3) \approx O(2mnp + m^3 p) + \sum_{i=1}^{K} O(m^2 p_i^2). \quad (20)$$

Finally, the overall computational complexity of the IFDDL algorithm is:

$$l\left(O(nm^2 p^\varepsilon) + O(2mnp + m^3 p) + \sum_{i=1}^{K} O(m^2 p_i^2)\right), \quad (21)$$

where l is the number of iterations.

Similarly, the overall computational complexity of the FDDL algorithm in [15] is approximately $l\left(O(nm^2 p^\varepsilon) + O(2mnp)\right)$.

As $m < p << n$, in both methods FDDL and IFDDL the majority of the learning time spends on the sparse coding stage. It means that our proposed method IFDDL has complexity similar to the FDDL.

4. EXPERIMENTS AND EVALUATION

To evaluate the effectiveness of our algorithm, we conducted experiments on popular benchmark data sets for face recognition and digit recognition. One problem in all our experiments is parameter selection. Due to the large number of hyper parameters in IFDDL model, instead of re-determining all the parameters, we keep $\lambda_1, \lambda_2, \gamma, w$ as in previous experiments [15], and tune μ using 10-fold cross validation.

To compare between approaches, we evaluate the accuracy rate in recognition. The accuracy rate is the ratio between the number samples in the dataset that is correctly recognized and total number of samples in the dataset:

$$Accuracy\ rate = \frac{\#\ correctly\ recognized\ samples}{\#\ samples}. \quad (22)$$

4.1 Face recognition

ORL database [26] contains 400 face images of 40 people, each image has a size of 112×92 pixels. The images were taken under different lighting conditions, facial expressions (open or closed eyes, smiling or not smiling) and facial details (glasses or not). For each person, we select randomly 6 images for training and remaining is for testing. We use random face feature descriptors by [27] in these experiments. Each face image is projected onto a 300-dimensional vector with a randomly generated matrix. The learned dictionary has 240 atoms, or 6 atoms in each sub-dictionary. Parameters in IFDDL found by experiments are as follows: λ_1=0.005, λ_2=0.005, γ=0.001, w=0.5, and μ=0.0001. We compare our algorithm with related and recently proposed methods. The experimental results are described in Table 1. As one can see from the table, our method improves the accuracy of the state-of-the-art models.

Extended Yale B database [28] consists of 2414 frontal-face images of 38 subjects. There are about 64 images for each person. The origin images were cropped to 192×168 pixels. We setup experiments like in [5]: extract random features by project-

ing images to 504-dimensional space, select randomly half of images for training and the remaining for testing. We use using random-face features. The learned dictionary consists of 570 atoms, which corresponds to 15 atoms per person. We found that the best result of this method was given by setting λ_1 =0.005, λ_2 =0.05, γ =0.005, w =0.5, and μ =0.0012. Table 2 summaries the recognition rates by our approach and several related approaches. One can see that our method gives better results than existing recently proposed methods.

Table 1. Recognition accuracy on ORL database.

Algorithms	Accuracy
KSVD [13]	95.6
D-KSVD [23]	93.6
LC-KSVD [2013]	95.6
IDL [6]	95.7
FDDL [15]	96.3
IFDDL	**97.1**

Table 2. Recognition results on Extended Yale B database.

Algorithms	Accuracy
KSVD [13]	93.1
D-KSVD [23]	94.1
SRC (15 person) [27]	80.5
LLC (70 local bases) [29]	90.7
LC-KSVD [5]	95.0
IDL [6]	95.7
FDDL [15]	97.6
IFDDL	**97.9**

AR database [30] contains over 4000 frontal images of 126 people. For each person, 26 images were taken in two separated Sessions. As in [27], we chose a subset consisting of 50 male subjects and 50 female subjects for 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 same condition from Session 2 were used for testing. We use using random-face features. The parameters are setting as lows: λ_1 =0.005, λ_2 =0.005, γ =0.001, w =0.5, and μ =0.0022. The recognition rates of algorithms in our experiments were reported in Table 3. The results show that our algorithm outperformed other state-of-the-art dictionary learning methods.

Table 3. Recognition results of AR database.

Algorithms	Accuracy
D-KSVD [23]	85.4
SRC [27]	88.8
LC-KSVD [5]	89.7
FDDL [15]	93.5
IFDDL	**93.7**

4.1 Digit recognition

For Digit recognition, USPS database [31] is used to evaluate our algorithm. USPS database consists of 7291 training images and 2007 testing images of size 16×16. We compare our algorithm to various algorithms which their results were also reported in [13]. In this experiment, original image 16×16 is directly used as the feature and the dictionary of each class has 90 atoms in FDDL with $\lambda_1 = \gamma$ =0.1, λ_2 =0.001, w =0.005, and μ =0.0022. The results are showed in Table 4.

Table 4. Accuracy rates of various methods on digit recognition.

Algorithms	Accuracy
IFDDL	**97.51**
FDDL [15]	96.46 (*)[1]
IDL [6]	93.85
SRSC	93.95
REC-L	93.17
REC-BL	95.62
SDL-G	93.33
SDL-D	96.46
DLSI	96.02
KNN	94.8
SVM-Gauss	95.8

As one can see from the table, although FDDL result is very impressive in comparison to others, our method gives a significant improvement over it.

5. CONCLUSION

This paper has presented a new approach for an improvement of the FDDL algorithm by combining the Fisher Discrimination Dictionary Learning (FDDL) method and the Incoherent Dictionary Learning (IDL) method. We have proposed to incorporate the incoherence between atoms which represent a specific object class to improve the discrimination capacity of the sub-dictionaries. The combination allows one to exploit the advantages of both methods and the discrimination capacity of the entire dictionary. The experiments have been conducted on well-known data sets in computer vision. The experimental results have shown that our approach outperformed others recently approaches in dictionary learning based classification tasks. For future work, our approach can be extended in several ways. We plan to investigate the sharing features approaches such as [23] to improve the power of dictionary in classification tasks. The learning process currently employed gradient method for optimization, which is often slow to converge to an

[1] The reason that FDDL gives a better result in our implementation than in original paper could be the way the sub-dictionaries were initialized. We randomly select data from training set and use them as initial dictionary rather than randomizing the whole dictionary matrices.

optimal solution in practice. Therefore, another future work is to find an efficient way to optimize the learning process of the model.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2011.17. We would like to thank Nguyen Quang Manh and Nguyen Duc Tuan for their useful contribution in this work.

REFERENCES

1. J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, Vol. 17, No. 1, pp. 53–69.
2. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non- local sparse models for image restoration," in *ICCV*, 2009.
3. R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proceedings of the Twenty- third Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.
4. M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, Vol. 13, No. 4, 2001, pp. 863–882.
5. Z. Jiang, Z. Lin, and L.S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2013.
6. T. Lin, S. Liu, and H. Zha, "Incoherent Dictionary Learning for Sparse Representation," in *ICPR*, 2012.
7. R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, "Self-Taught Learning: Transfer Learning from Unlabeled Data," in *ICML*, 2007.
8. I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *CVPR*, 2010.
9. F. Rodriguez and G. Sapiro, "Sparse representation for image classification: Learning discriminative and reconstructive non-parametric dictionaries," in *Technical report, DTIC Document*, 2008.
10. S. G. Mallat, and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, Vol. 41, No. 12, 1993, pp. 3397-3415.
11. Y.C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40-44.
12. K. Engan, S. Aase, and J. Husoy, "Frame based signal compression using method of optimal directions (MOD)," in *Proc. of IEEE Int. Symp. Circuits and Systems*, 1999.
13. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transaction on Signal Processing*, Vol. 54, No. 11, 2006, pp. 4311–4322.

14. Q. Zhang and B.X. Li, “Discriminative K-SVD for dictionary learning in face recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
15. M. Yang, L. Zhang, X. Feng, and D. Zhang, “Fisher Discrimination Dictionary Learning for Sparse Representation,” in *ICCV*, 2011.
16. M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transaction on Image Processing*, Vol. 15, No. 12, 2006, pp. 3736–3745.
17. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zissserman, “Discriminative learned dictionaries for local image analysis,” in *CVPR*, 2008.
18. D. Pham and S. Venkatesh, “Joint learning and dictionary construction for pattern recognition,” in *Proceedings of CVPR*, 2008.
19. J. C. Yang, K. Yu, and T. Huang, “Supervised Translation-Invariant Sparse coding,” in *CVPR*, 2010.
20. M. Yang, L. Zhang, J. Yang and D. Zhang, „Metaface learning for sparse representation based face recognition,” in *ICIP*, 2010.
21. N. Zhou, Y. Shen, J. Peng, and J. Fan, "Learning inter-related visual dictionary for object recognition," In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3490-3497.
22. S. Kong and D. Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in *ECCV*, 2012, pp. 186-199.
23. H. Wang, H. Zhou, and A. Finn, "Discriminative dictionary learning via shared latent structure for object recognition and activity recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6299-6304.
24. M. Yang, D. Dengxin, L. Shen, and L.V. Gool, "Latent dictionary learning for sparse representation based classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4138-4145.
25. S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “A interior-point method for large-scale l1-regularized least squares,” in *IEEE Journal on Selected Topics in Signal Processing* 1, 2007, pp. 606–617.
26. F. S. Samaria and A.C. Harter, “Parameterization of a stochastic model for human face identification,” in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.
27. J. Wright, A. Yang, A. Ganesh, S. Satry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE-Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.
28. A. S. Georghiades, P.N. Belhumeur, D.J. Kriegman, “From few to many: illumination cone models for face recognition under variable lighting and pose,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.
29. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality Constrained Linear Coding for Image Classification,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
30. A. Martinez and R. Benavente, “The AR Face Database,” in *CVC Technical Report 24*.
31. USPS Handwritten Digit Database. Last accessed August 15 2015. http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html.

Thuy Thi Nguyen received her Ph.D. degree in Computer Science from Graz University of Technology, Austria, in 2009. She is currently lecturer, researcher and head of Department of Computer Science, Faculty of Information Technology,Vietnam National University of Agriculture. Her research interests include computer vision, machine learning, pattern recognition, and applications.

Binh Thanh Huynh is an Associate Professor at Computer Science Department, School of Information and Communication Technology (SoICT), Hanoi University of Science and Technology (HUST), Vietnam. Her research area: Computational Intelligence - genetic algorithms, heuristic algorithms. She is Chair of IEEE Vietnam Computational Intelligent Society (IEEE Vietnam CIS). She is reviewer for more than 20 international journals and conferences. More information can be found at: http://www.soict.hust.edu.vn/~binhht/.

Sang Viet Dinh received the Ph.D. degree in Computer Science from Dorodnitsyn Computing Centre of Russian Academy of Sciences (CCRAS) in 2013. He is currently working at Computer Science Department, School of Information and Communication Technology (SoICT), Hanoi University of Science and Technology (HUST), Vietnam. His research interests include discrete mathematics, machine learning and computer vision.