Label Associated Dictionary Pair Learning for Face Recognition

Dao Duy Son
Hanoi University of Science and Technology
sondaoduy@gmail.com

Dinh Viet Sang*
Hanoi University of Science and Technology
sangdv@soict.hust.edu.vn

Huynh Thi Thanh Binh
Hanoi University of Science and Technology
binhht@soict.hust.edu.vn

Nguyen Thi Thuy
Faculty of Information Technology
Vietnam National University of Agriculture
ntthuy@vnua.edu.vn

ABSTRACT

Dictionary learning (DL) has been successfully applied to various pattern classification tasks. Sparse coding has played a vital role in the success of such DL-based models. However, the popular sparsity constraints using l_0 or l_1-norm often make the training phase time-consuming. Recently, an emerging trend in using l_2-norm has shown its advantages in both accuracy and computational speed. However, the supervised approach that exploits label information in the training phase has not been investigated in such l_2-norm based methods. In this paper, we propose a novel supervised dictionary learning method that incorporates label information in the objective function. Based on that, we also propose an effective classification schema. Experiments on three popular face recognition datasets show that our method has promising results. Especially, our method has extremely fast speed in test phase, while maintaining competitive accuracy in comparison with other state-of-the-art models.

CCS Concepts

•Computing methodologies → Supervised learning by classification; Regularization;

Keywords

Dictionary Learning; Sparse Coding; Supervised Learning; Pattern Recognition; Face Recognition

1. INTRODUCTION

Sparse coding has proven to be an powerful tool for many applications such as image denoising [10, 11], audio processing [5, 23] and, especially in image classification [8, 13, 17, 19, 22]. The purpose of sparse coding is to approximate an input signal by a linear combination of some atoms in an over-complete dictionary. Sparse coding problem can be expressed as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|x - \mathbf{D}\alpha\|_2^2 \quad s.t.\ \|\alpha\|_0 \leq T, \tag{1}$$

where $x \in \mathbb{R}^m$ is an original input signal, $\mathbf{D} \in \mathbb{R}^{m \times p}$ is the dictionary consisting of p atoms, $\alpha \in \mathbb{R}^p$ is the sparse code and T is the sparsity level which limits the number of nonzero elements in the sparse code. In most cases, the dictionary is overcomplete, i.e. $p \gg m$. Sparse representation converts the low dimensional space of input signals into a higher dimensional space of spare codes, which can be more easily separated and, therefore, enhance the classification efficiency. Based on the formulation of sparse coding in Eq. (1), many methods have successfully exploited supervised paradigm that supplements the traditional dictionary learning process with a label-associated term in order to improve classification accuracy [8, 22].

In general, sparse coding is an NP-hard problem [3]. Many approximate optimization techniques have been studied to solve the sparse coding problem, but they often require high computational complexity. For instance, OMP [14] is one of the most popular algorithms for finding the sparse code. The high cost computation is a huge constraint of sparse code in some applications that require fast computing for efficient classification. Recently, some methods such as [6,20,21] have demonstrated impressive results by using l_2-norm instead of l_0 and l_1-norm. Nevertheless, in those methods, the label information of training samples has not been investigated in learning dictionaries.

In this paper, we present a supervised dictionary learning method that integrates conventional dictionary pair learning model with a label-associated term. Our proposed method shows promising result in popular datasets for human face recognition such as Extended YaleB, AR and JAFFE. The proposed method achieves extremely fast speed in test phase, while maintaining competitive classification accuracy.

*Corresponding author.

SoICT '16, December 08-09, 2016, Ho Chi Minh City, Viet Nam

DOI: http://dx.doi.org/10.1145/3011077.3011105

The paper is organized as follows: In section 1 and 2, we quickly introduce about dictionary learning and related work on sparse dictionary learning. In section 3, we present the main idea of supervised dictionary learning, projective dictionary pair learning and our proposed method one after the other in subsections. Optimization, experiments and evaluation are described in the successive subsections. Conclusion is in section 4 with some discussions about future work.

2. RELATED WORK

Learning an overcomplete dictionary plays a critical role in the success of many sparse coding based classification methods in various applications. A desired dictionary should faithfully represent the query samples while supporting the discrimination of object classes. Based on KSVD [1] Zhang and Li [22] proposed a joint learning algorithm DKSVD that incorporates classification error into the objective function of KSVD. Jiang et al. [8] proposed a new label consistence constraint and combined it with the reconstruction and classification errors to form a new algorithm called LC-KSVD. In addition to enforcing discriminative constraints on the dictionary, Yang et al. [19] proposed a method called FDDL that uses Fisher criterion to make the sparse codes more discriminative.

Typically, in most of the existing state-of-the-art DL methods, the standard l_0 or l_1-norm sparsity constraint is used on the representation coefficients. However, these kinds of constraint usually lead to time-consuming in training and testing phases. In order to avoid expensive computation cost in using l_0 or l_1-norm constraints, Gu et al. in [6] proposed a dictionary learning method called DPL that uses l_2-norm constraint. DPL [6] learns jointly a pair of separate dictionaries: a synthesis dictionary for representation power and an analysis dictionary for classification power. Go further in enhancing the classification accuracy, Shuang Yuan et al [20] incorporated Fisher criterion into the DPL objective function to improve the discrimination of the pair of dictionaries. Compared with conventional DL methods, l_2-norm based DL approach not only significantly reduces time complexity but also leads to very competitive results in image classification tasks.

Our proposed method focuses not only on the classification accuracy but also aims to elevate the classification time by integrating the projective dictionary pair learning with a label-associated term. The proposed method is significantly faster while having competitive classification accuracy in comparison with the DPL method when performing experiments on popular face recognition datasets.

3. DISCRIMINATIVE DICTIONARY LEARNING WITH LABEL INFORMATION

3.1 Discriminative sparse coding and supervised dictionary learning

In general, sparse codes are used directly as features to train a classifier for classification task. This idea has been followed by a lot of methods such as [2, 7, 15, 18]). However, the sparse codes in this situation are not discriminative enough for some complex classification tasks. To enhance the discrimination of sparse codes, many methods such as D-KSVD [22] combine the dictionary learning with classifier training into one common objective function.

Suppose that $\mathbf{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}$ is the set of n training samples $x_i \in \mathbb{R}^m$ from K classes. D-KSVD method learns a dictionary with a classifier concurrently as follows:

$$\langle \mathbf{D}^*, \mathbf{W}^*, \mathbf{A}^* \rangle = \underset{\mathbf{D},\mathbf{W},\mathbf{A}}{\operatorname{argmin}} \left(\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \beta \|\mathbf{H} - \mathbf{W}\mathbf{A}\|_F^2 + \gamma \|\mathbf{W}\|_F^2 \right) \quad s.t. \forall i, \|a_i\|_0 \le T, \|d_i\|_2^2 = 1, \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{m\times p}$ is the dictionary, $\mathbf{A} \in \mathbb{R}^{p\times n}$ is matrix of sparse codes, $\mathbf{W} \in \mathbb{R}^{K\times p}$ is the multi-class linear classifier, $\mathbf{H} = \left[h_1, h_2, \ldots, h_n\right] \in \mathbb{R}^{K\times n}$ is the binary label matrix of n training samples. Here $h_j = \left[0, 0, \cdots, 1, \cdots, 0\right]^T \in \mathbb{R}^K$ is the one-hot coding vector of sample x_j, i.e. the position of nonzero element indicates the class that x_j belong to. The discrimination term $\|\mathbf{H} - \mathbf{W}\mathbf{A}\|_F^2$ is the linear prediction error produced by $\mathbf{W}$ and $\mathbf{A}$, which is used to make $\mathbf{A}$ more discriminative. The regularization term $\|\mathbf{W}\|_F^2$ is used to prevent elements of $\mathbf{W}$ from getting high values, which would often lead to overfitting.

LC-KSVD goes further to make the sparse code more discriminative. The method integrates a label consistent term into the objective function as follows:

$$\langle \mathbf{D}^*, \mathbf{M}^*, \mathbf{W}^*, \mathbf{A}^* \rangle = \underset{\mathbf{D},\mathbf{M},\mathbf{W},\mathbf{A}}{\operatorname{argmin}} \left(\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{M}\mathbf{A}\|_F^2 + \beta \|\mathbf{H} - \mathbf{W}\mathbf{A}\|_F^2 + \gamma \|\mathbf{W}\|_F^2 \right) \quad s.t.\ \forall i, \|a_i\|_0 \le T, \|d_i\|_2^2 = 1, \tag{3}$$

where $\|\mathbf{Q} - \mathbf{M}\mathbf{A}\|_F^2$ is the label consistent term, $\mathbf{M}$ is a transformation matrix that converts the sparse code matrix $\mathbf{A}$ into a predefined discriminative form $\mathbf{Q} = \left[q_1, q_2, \ldots, q_n\right] \in \mathbb{R}^{K\times n}$, and $q_i \in \mathbb{R}^K$ is a binary vector, where the non-zero values appear at those entries where the sample x_i and the dictionary atom d_k share the same label. For instance, assuming that $\mathbf{X} = \left[x_1, x_2, x_3, x_4\right], \mathbf{D} = \left[d_1, d_2, d_3, d_4\right]$, where x_1, d_1, d_2 from the first class and x_2, x_3, x_4, d_3, d_4 from the second class, then the corresponding matrix $\mathbf{Q}$ can be expressed as follows:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

Both two equations (2) and (3) can be solved by applying K-SVD algorithm with some renormalization strategies on $\mathbf{D}$ and $\mathbf{W}$.

3.2 Dictionary Pair Learning (DPL)

Suppose that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$ is the set of training samples, where $\mathbf{X}_k$ is the set of training samples from class k. The aim of training DPL is to learn a synthesis dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K]$ and an analysis dictionary $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K]$ to reconstruct input samples. The synthesis dictionary $\mathbf{D}$ serves the representation task and the analysis dictionary $\mathbf{P}$ serves the classification task. A pair $\{\mathbf{D}_k \in \mathbb{R}^{m\times t}$ and $\mathbf{P}_k \in \mathbb{R}^{t\times m}\}$ is corresponded subdictionaries for class k, where t is the number of atoms in each subdictionary $\mathbf{D}_k$.

The objective function of DPL proposed in [6] is represented as follows:

$$\langle \mathbf{D}^*, \mathbf{P}^* \rangle = \underset{\mathbf{D},\mathbf{P}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(\|\mathbf{X}_k - \mathbf{D}_k\mathbf{P}_k\mathbf{X}_k\|_F^2 + \beta \|\mathbf{P}_k\overline{\mathbf{X}_k}\|_F^2 \right) \quad s.t.\ \forall i, \|d_i\|_2^2 \leq 1, \tag{4}$$

where $\beta > 0$ is a constant.

The analysis dictionary $\mathbf{P}$ is constructed by analytically computing the sparse codes $\mathbf{A} = \mathbf{P}\mathbf{X}$. Matrix $\overline{\mathbf{X}}_k$ is a complementary matrix of $\mathbf{X}_k$ from the whole training data $\mathbf{X}$. The term $|\mathbf{P}_k\overline{\mathbf{X}_k}\|_F^2$ is used to enforce the discrimination of the DPL model. Based on the studies on sparse subspace clustering, the authors design the analysis dictionary $\mathbf{P}$ such that the subdictionary $\mathbf{P}_k$ only effectively projects the samples from class k, i.e. $\mathbf{P}_k\mathbf{X}_i \approx 0, \forall k \neq i$. On the other hand, the subdictionary $\mathbf{D}_k$ with the projective code $\mathbf{P}_k\mathbf{X}_k$ should well reconstruct the signal $\mathbf{X}_k$. Therefore, it comes to a conclusion that the pair of dictionaries should minimize the reconstruction error:

$$\sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{D}_k\mathbf{P}_k\mathbf{X}_k\|_F^2 \rightarrow min \tag{5}$$

For a testing sample y from class k, its projective code $\mathbf{P}_k y$ must be bigger than others $\mathbf{P}_i y, \forall i \neq k$, so that the residual $\|y - \mathbf{D}_k\mathbf{P}_k y\|_2^2$ should be much smaller than the residuals $\|y - \mathbf{D}_i\mathbf{P}_i y\|_2^2, \forall i \neq k$. The formula for classification task associated with DPL is:

$$\text{identity(y)} = \underset{k}{\operatorname{argmin}} \|y - \mathbf{D}_k\mathbf{P}_k y\|_2^2 \tag{6}$$

3.3 Label Associated Dictionary Pair Learning(LADPL)

Note that DPL gives state-of-the-art performance on certain image classification tasks. Inspired by the idea of using analysis dictionary $\mathbf{P}$ for classification in DPL, and aiming at exploiting label information in learning DPL, we propose to combine the pair dictionary learning with a classifier training term. This makes the classification task simpler as it depends just on $\mathbf{P}$ and the classifier. The objective function is constructed as follows:

$$\langle \mathbf{D}^*, \mathbf{W}^*, \mathbf{P}^* \rangle = \underset{\mathbf{D},\mathbf{W},\mathbf{P}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(\|\mathbf{X}_k - \mathbf{D}_k\mathbf{P}_k\mathbf{X}_k\|_F^2 + \beta \|\mathbf{P}_k\overline{\mathbf{X}_k}\|_F^2 + \gamma \|\mathbf{H}_k - \mathbf{W}_k\mathbf{P}_k\mathbf{X}_k\|_F^2 \right) s.t.\ \|d_i\|_2^2 \leq 1, \forall i \tag{7}$$

where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$ is the set of training samples with $\mathbf{X}_k$ is the set of training samples from class k. $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K]$ and $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K]$ are synthesis dictionary and analysis dictionary. The synthesis dictionary serves the representation task and the analysis dictionary serves the classification task. A pair $\{\mathbf{D}_k \in \mathbb{R}^{m \times t}$ and $\mathbf{P}_k \in \mathbb{R}^{t \times m}\}$ is corresponding to subdictionary for class k, where t is the number of atoms in each subdictionary $\mathbf{D}_k$. $\mathbf{H} = \left[\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\right] \in \mathbb{R}^{K \times n}$ is the binary label matrix of n training samples with $\mathbf{H}_k$ is the label matrix of class k which has all elements in row k-th equal 1 and the other elements equal 0. $\mathbf{W} = \left[\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K\right]$ is the classifier where each element $\mathbf{W}_k \in \mathbb{R}^{K \times t}$ is corresponded for the data from class k. β and γ are scalars.

3.4 Optimization for LADPL

Algorithm 1 Label Associated Dictionary Pair Learning

Input:
Training samples $\mathbf{X} \in \mathbb{R}^{m \times n}$
Label Matrix $\mathbf{H} \in \mathbb{R}^{K \times n}$
Subdictionary size t
Weighting parameters $\alpha, \beta, \gamma, \lambda$

Output:
Synthesis dictionary $\mathbf{D}$, analysis dictionary $\mathbf{P}$ and classifier $\mathbf{W}$

```
Initialize the dictionaries D, P and classifier W by random matrices of size m×(K×t), (K×t)×m and m×n respectively. Each column of D, P and W is then normalized to have unit l2−norm.
while not converge do
    for i = 1 to K do
        Fix D, P,W and update A using (10);
        Fix A and update D using (14);
        Fix A and update P using (12);
        Fix A and update W using (15);
    end for
end while
```

Similar to the construction of DPL, we introduce a new matrix $\mathbf{A}$ to relax the problem (7) as follows:

$$\langle \mathbf{D}^*, \mathbf{W}^*, \mathbf{P}^*, \mathbf{A}^* \rangle = \underset{\mathbf{D},\mathbf{W},\mathbf{P},\mathbf{A}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(\|\mathbf{X}_k - \mathbf{D}_k\mathbf{A}_k\|_F^2 + \alpha \|\mathbf{P}_k\mathbf{X}_k - \mathbf{A}_k\|_F^2 + \beta \|\mathbf{P}_k\overline{\mathbf{X}_k}\|_F^2 + \gamma \|\mathbf{H}_k - \mathbf{W}_k\mathbf{A}_k\|_F^2 \right) \ s.t.\ \forall i, \|d_i\|_2^2 \leq 1. \tag{8}$$

where α, β, γ are scalar constants.

3.4.1 Initialization

First of all, we initialize the synthesis dictionary $\mathbf{D}$, analysis dictionary $\mathbf{P}$ and classifier $\mathbf{W}$ as random matrices and then normalize each atom of them to have unit l_2-norm. After that, we update $\mathbf{A}$ and $\mathbf{D}$,$\mathbf{P}$,$\mathbf{W}$ respectively.

3.4.2 Fix $\mathbf{D}$, $\mathbf{P}$ and $\mathbf{W}$ and update $\mathbf{A}$

The rule for updating $\mathbf{A}$ can be expressed as follows:

$$\mathbf{A}^* = \underset{\mathbf{A}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(\|\mathbf{X}_k - \mathbf{D}_k\mathbf{A}_k\|_F^2 + \alpha \|\mathbf{P}_k\mathbf{X}_k - \mathbf{A}_k\|_F^2 + \gamma \|\mathbf{H}_k - \mathbf{W}_k\mathbf{A}_k\|_F^2 \right) \tag{9}$$

Problem (9) can be analytically solved by using least square method:

$$\mathbf{A}_k^* = (\mathbf{D}_k^T\mathbf{D}_k + \alpha\mathbf{I} + \gamma\mathbf{W}_k^T\mathbf{W}_k)^{-1}(\alpha\mathbf{P}_k\mathbf{X}_k + \mathbf{D}_k^T\mathbf{X}_k + \gamma\mathbf{W}_k^T\mathbf{H}_k) \tag{10}$$

3.4.3 Fix $\mathbf{A}$ and update $\mathbf{P}$, $\mathbf{D}$, $\mathbf{W}$

The rule for updating $\mathbf{P}$, $\mathbf{D}$ and $\mathbf{W}$ can be expressed as

follows:

$$\begin{cases} \mathbf{P}^* = \underset{\mathbf{P}}{\operatorname{argmin}} \sum_{k=1}^{K} (\alpha \|\mathbf{P}_k \mathbf{X}_k - \mathbf{A}_k\|_F^2 + \beta \|\mathbf{P}_k \overline{\mathbf{X}_k}\|_F^2) \\ \mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{D}_k \mathbf{A}_k\|_F^2 \ s.t. \ \forall i, \|d_i\|_2^2 \leq 1 \\ \mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{k=1}^{K} \gamma \|\mathbf{H}_k - \mathbf{W}_k \mathbf{A}_k\|_F^2 \end{cases} \quad (11)$$

The closed-form solution of $\mathbf{P}^*$ is:

$$\mathbf{P}_k^* = \alpha \mathbf{A}_k \mathbf{X}_k^T (\alpha \mathbf{X}_k \mathbf{X}_k^T + \beta \overline{\mathbf{X}}_k \overline{\mathbf{X}}_k^T + \theta \mathbf{I})^{-1}. \quad (12)$$

where $\theta = 10\mathbf{e}^{-4}$ is a small number. The $\mathbf{D}$ problem can be optimized by introducing a variable $\mathbf{S}$:

$$\min_{\mathbf{D},\mathbf{S}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{D}_k \mathbf{A}_k\|_F^2 \ s.t. \ \mathbf{D} = \mathbf{S}, \|s_i\|_2^2 \leq 1. \quad (13)$$

We can apply the ADMM algorithm to optimize the problem (13) as follows:

$$\begin{cases} \mathbf{D}^{(r+1)} = \underset{\mathbf{D}}{\operatorname{argmin}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{D}_k \mathbf{A}_k\|_F^2 + \rho \|\mathbf{D}_k - \mathbf{S}_k^{(r)} + \mathbf{T}_k^{(r)}\|_F^2 \\ \mathbf{S}^{(r+1)} = \underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \rho \|\mathbf{D}_k^{(r+1)} - \mathbf{S}_k + \mathbf{T}_k^{(r)}\|_F^2 \quad s.t. \|s_i\|_2^2 \leq 1 \\ \mathbf{T}^{(r+1)} = \mathbf{T}^{(r)} + \mathbf{D}_k^{(r+1)} - \mathbf{S}_k^{(r+1)}, \text{update } \rho \text{ if appropriate.} \end{cases} \quad (14)$$

We can update the classifer $\mathbf{W}_k^*$ by using least-square method:

$$\mathbf{W}_k^* = (\gamma \mathbf{A}_k \mathbf{A}_k^T)^{-1} \gamma \mathbf{H}_k \mathbf{A}_k^T \quad (15)$$

In experiments, we update $\mathbf{D}$,$\mathbf{P}$,$\mathbf{W}$ and $\mathbf{A}$ until the difference between the values of objective function in Eq. (8) in two consecutive iteration is less than 0.01. The outcome of the process is the synthesis dictionary $\mathbf{D}$, the analysis dictionary $\mathbf{P}$ and a classifier $\mathbf{W}$. The pseudo-code of overall optimization procedure is described in the Algorithm 1.

3.4.4 Classification Model

With the proposed method for learning a pair dictionary in combination with label information, we can now build a simple yet effective model for classification. For a testing sample y, we first calculate $v = \mathbf{W}\mathbf{P}y$, where $\mathbf{W}$ and $\mathbf{P}$ are the outputs of training algorithm 1, and $v = [v_1, v_2, \ldots, v_K] \in \mathbb{R}^K$. Then the position of the largest element in v will indicate the predictive label for class of testing sample y.

$$\text{identity(y)} = \underset{k}{\operatorname{argmax}} \ v_k. \quad (16)$$

The advantage of this classification scheme is it directly uses the testing input signal with the analysis dictionary and the learnt classifier for recognition. It makes the testing phase of LADPL is extremely faster than the original version of DPL and other state-of-the-art models. Note that the DPL model has to calculate the residual function $\|y - \mathbf{D}_k \mathbf{P}_k y\|_2^2$ by number of class times. This means the more number of classes the dataset has, the more time it needs to take to classify a testing sample.

3.4.5 Convergence Analysis

To optimize the objective function in (7), we alternatively optimize $\mathbf{A}$ and $\mathbf{D}, \mathbf{P}, \mathbf{W}$. The experiments show that the objective function always decreases after every iteration.

Fig.1 shows the convergence curve of our algorithm. In practice, our algorithm often converges after 30 iterations.

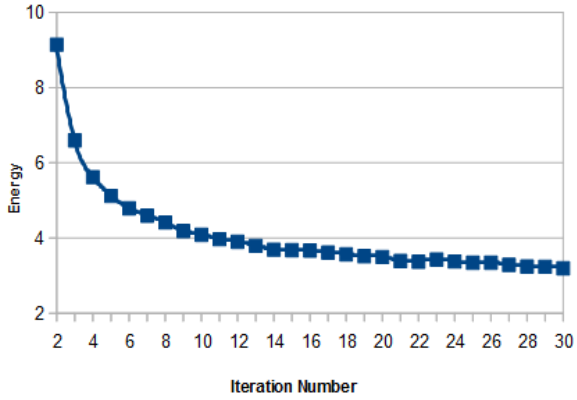

Figure 1: The convergence curve of LADPL on Extended YaleB.

3.5 Experimental Results

We perform a set of experiments to evaluate our proposed method. We apply our method on three popular datasets for face recognition to evaluate the performance in terms of classification accuracy and time consuming in testing phase.

For a convenience in implementation and evaluation, we implemented our proposed algorithm in MatLab. The configuration of the computer used in our experiments is described as in Table 1.

Table 1: Computer configuration

Hardware	Configuration
CPU	Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz 2.40GHz
RAM	8.00 GB
Disk	500 GB

3.5.1 Dataset

In our experiments, we use two popular benchmark datasets for face recognition (Extended YaleB [4] and AR [12]) and one data set for facial expression recognition, the Japanese Female Facial Expression (JAFFE) dataset [9]. We use the same features for face representation as in [8] for Extended YaleB and AR. This is to guarantee a fair comparison with other methods. For the JAFFE data set, random features from the raw images are used.

The Extended YaleB database has large differences in illumination and expressions. The AR database contains many differences of illumination, expressions and sunglass and scarf occlusion. The JAFFE includes images of emotional expression.

We use random-face features as [16] to represent image samples in the datasets. Each raw picture in the dataset will be projected into a new space by a random matrix generated from a zero-mean normal distribution. In Extended YaleB and JAFFE datasets, the dimension of the new spaces is 504, while in the AR dataset, the dimension of the new space is 540.

Figure 2: Example of images from Extend YaleB dataset.

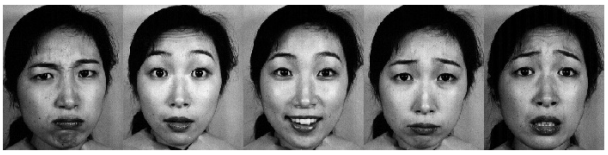

Figure 3: Example of images in JAFFE dataset.

3.5.2 Results

In the proposed model as well as in related works, there are meta parameters to tune in order to get the best performance of the algorithms. They are dictionary size, α, β, γ. After extensive experiments, we found that with the following parameters setting we can get the best performance for our model. These parameters must be experimented and set differently for each dataset. Random search is used to estimate the best values of parameters, which are shown in Table 2.

Table 2: Parameter setting for our model for each data set

Dataset	SubDictionary size	α	β	γ	θ
Extended YaleB	30	0.019	0.003	0.0135	0.0001
AR	30	0.02	0.003	0.015	0.0001
JAFFE	30	0.03	0.005	0.0135	0.0001

We compare our method with DPL and LC-KSVD on three aspects: accuracy, training time and testing time. The results for LC-KSVD was taken from the paper [6]. We also implemented DPL method with the computer configuration as in Table 1 to evaluate its performance efficiency.

We calculate the accuracy of classification by taking the ratio of true predicted samples over total number of testing samples.

Firstly with Extended YaleB, the dataset includes 2414 face images of 38 people. We choose randomly half of the dataset for training and half for testing. The feature dimension is 504 as we use random-face features provided by Jiang et al [8]. The results are shown in Table 3.

Table 3: Experiment results on Extended YaleB

Method	Accuracy (%)	Training time (s)	Testing time (s)
LC-KSVD	96.7	412.58	4.22e-4
DPL	97.6	5.916	2.37e-4
LADPL	96.9	7.557	4.17e-6

With AR, the dataset includes 2600 pictures of 50 females, 50 males, each person has 26 pictures. We choose 20 for training and 6 for testing randomly. The feature dimension is 540 as we use random-face features provided by Jiang et al [8]. The results are shown in Table 4.

Table 4: Experiment results on AR

Method	Accuracy (%)	Training time (s)	Testing time (s)
LC-KSVD	97.8	1,806.3	7.72e-4
DPL	98.9	17.509	6.5e-4
LADPL	99.0	24.696	0.3e-4

Finally with JAFFE, the dataset includes 213 images of 10 different Japanese women. Each woman has 7 different emotional expressions. We use 3/4 of each emotion for training and 1/4 of each emotion for testing (about 164 images for training and 49 image for testing). The feature dimension is 504 as we use random-face features provided by Wright et al [16]. The results are shown in Table 5.

Table 5: Experiment results on JAFFE

Method	Accuracy (%)	Training time (s)	Testing time (s)
DPL	95.9	0.965	4.28e-5
LADPL	95.7	1.436	1.02e-5

From the experimental results as shown in the above Tables 3-5 we can see that our proposal method can give state-of-the-art results in term of classification accuracy. In term of training time, our model takes a reasonable time for learning. This learning time is longer compared to DPL. However, it is more important to have a model with less computational time for testing. For our proposed model, in term of testing time we can get from 4 upto 50 times faster than the state-of-the-art DPL model. This is a great advantage of our model in comparison with related models.

4. CONCLUSION

We have presented an effective method for learning dictionary for spare representation using a pair of synthesis and analysis dictionaries. By incorporating classifier training into the projective dictionary pair learning we can get a discriminative dictionary for representation meanwhile improving classification performance. Our experimental results have shown advantages of our proposed model by reducing a huge amount of testing time while maintaining state-of-the-art classification accuracy. For future work, we will investigate further constraints on learning subdictionary and learning a shared dictionary.

5. ACKNOWLEDGMENTS

This research is funded by Hanoi University of Science and Technology under grant number T2015-221.

6. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR),*

2010 IEEE Conference on, pages 2559–2566. IEEE, 2010.

[3] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[4] A. S. Georghiades, P. N. Belhumeur, and 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*, 23(6):643–660, 2001.

[5] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng. Shift-invariance sparse coding for audio classification. *arXiv preprint arXiv:1206.5241*, 2012.

[6] S. Gu, L. Zhang, W. Zuo, and X. Feng. Projective dictionary pair learning for pattern classification. In *Advances in Neural Information Processing Systems*, pages 793–801, 2014.

[7] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in neural information processing systems*, pages 609–616, 2006.

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

[9] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database, 1998.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009.

[11] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.

[12] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[13] T. T. Nguyen, T. T. B. Huynh, and V. S. Dinh. A new approach for learning discriminative dictionary for pattern classification. *Journal of Information Science and Engineering*, 32(4):1113–1127, 2016.

[14] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

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

[16] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

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

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

[19] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *2011 International Conference on Computer Vision*, pages 543–550. IEEE, 2011.

[20] S. Yuan, H. Zheng, and D. Lin. Image classification based on discriminative dictionary pair learning. In *Chinese Conference on Biometric Recognition*, pages 176–185. Springer, 2015.

[21] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *2011 International Conference on Computer Vision*, pages 471–478. IEEE, 2011.

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

[23] M. Zibulevsky and B. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.