Pattern Recognition 46 (2013) 2134–2143

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Joint discriminative dimensionality reduction and dictionary learning for face recognition

Zhizhao Feng [1], Meng Yang [1], Lei Zhang *, Yan Liu, David Zhang

*Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*

ARTICLE INFO

*Article history:*
Received 3 February 2012
Received in revised form
10 January 2013
Accepted 12 January 2013
Available online 24 January 2013

*Keywords:*
Dictionary learning
Face recognition
Dimensionality reduction
Collaborative representation

ABSTRACT

In linear representation based face recognition (FR), it is expected that a discriminative dictionary can be learned from the training samples so that the query sample can be better represented for classification. On the other hand, dimensionality reduction is also an important issue for FR. It cannot only reduce significantly the storage space of face images, but also enhance the discrimination of face feature. Existing methods mostly perform dimensionality reduction and dictionary learning separately, which may not fully exploit the discriminative information in the training samples. In this paper, we propose to learn jointly the projection matrix for dimensionality reduction and the discriminative dictionary for face representation. The joint learning makes the learned projection and dictionary better fit with each other so that a more effective face classification can be obtained. The proposed algorithm is evaluated on benchmark face databases in comparison with existing linear representation based methods, and the results show that the joint learning improves the FR rate, particularly when the number of training samples per class is small.



## 1. Introduction

Face recognition (FR) methods have been studied for over 30 years, and various techniques have been developed [1–8,13] to handle different problems in face recognition, such as illumination, pose, occlusion and small sample size, etc. Face images usually have a high dimensionality, which makes the storage space high and increases the computational cost. In addition, the high dimensionality also decreases the discrimination of face images. Therefore, many dimensionality reduction techniques [2,9,10–12,14] have been developed to reduce the dimension of face images and enhance the discriminative features. The representative dimensionality reduction methods include Principal Component Analysis (PCA) [9], Linear Discriminate Analysis (LDA) [10], Locality Preserving Projection (LPP) [2], etc. These so-called subspace analysis based FR methods are simple to apply; however, they are less effective to handle the expression and illumination changes. When the training samples are insufficient, the subspace learned by these methods will be much biased.

In the subspace based FR methods, often the nearest neighbor (NN) classifier and SVM are used for the classification. Recently, a new face classification scheme, i.e., the sparse representation based classification (SRC) [6], was proposed. In SRC, a query face image is encoded over the original training set with sparsity constraint imposed on the encoding vector. The training set acts as a dictionary to represent the testing samples as a sparse linear combination of its atoms. The classification is then performed by checking which class leads to the smallest reconstruction residual of the query sample. The SRC classifier shows very competitive performance, but its performance will drop much when the training samples per class are insufficient. It is also claimed in [6] that dimensionality reduction is no longer critical in the SRC scheme and random projection can achieve similar results to PCA and LDA when the dimensionality is high enough. Nonetheless, if a lower dimensionality is required, PCA and LDA will have clear advantage over random projection. Some works [14,15] has been done to investigate the dimensionality reduction for SRC. For example, Zhang et al. [14] proposed an unsupervised learning method for dimensionality reduction in SRC, and it leads to higher FR rates than PCA and random projection. This validates that a well designed dimensionality reduction method can benefit the sparse classification scheme.

In the meantime, there has been an increasing interest in learning a dictionary to represent the query image instead of using the original training samples. In FR, the original face images may contain some redundant information, noise or other trivial information that will obstruct the correct recognition. In [16], Yang et al. proposed a metaface learning (MFL) algorithm to represent the training samples by a series of "metafaces" learnt

* Corresponding author. Tel: 852 2766 7355.
*E-mail address:* cslzhang@comp.polyu.edu.hk (L. Zhang).
[1] The first two authors contribute equally to this work.

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.016

from each class. Based on the classical KSVD algorithm [17], in [18] a DKSVD algorithm was developed to code the query image and use the coding coefficients for classification. In [19], a supervised algorithm was proposed to learn a dictionary as well as a classifier for image classification tasks (e.g., digit recognition, texture classification). In [20], a class-dependent supervised simultaneous orthogonal matching pursuit scheme was developed to solve the dictionary learning problem while increasing the inter-class discrimination. Very recently, a Fisher discrimination dictionary learning algorithm [3] was developed for sparse representation based pattern classification, and it shows very competitive performance with other dictionary learning based pattern classification schemes.

The dimensionality reduction (DR) and dictionary learning (DL) are mostly studied as two independent problems in FR. Usually, DR is performed first to the training samples and the dimensionality reduced data are used for DL. However, the pre-learned DR projection may not preserve the best features for DL. Intuitively, the DR and DL processes should be jointly conducted for a more effective FR. To this end, we propose a joint discriminative DR and DL (JDDRDL) scheme to exploit more effectively and robustly the discriminative information of training samples. The goal is that the face image features from different classes can be effectively separated by a dictionary in a subspace, which are to be determined. In the proposed JDDRDL, an energy functional is defined and an iterative optimization algorithm is given to alternatively optimize the dictionary and projection matrix. From some initialization, in each iteration, for a fixed projection $\boldsymbol{P}$, the desired dictionary $\boldsymbol{D}$ can be updated; then with the updated dictionary $\boldsymbol{D}$, the projection matrix $\boldsymbol{P}$ can be refined. After several iterations, the learned $\boldsymbol{P}$ and $\boldsymbol{D}$ together can lead to a more effective FR system.

One important advantage of the proposed JDDRDL scheme is that it is more robust to the small sample size problem than state-of-art linear representation based face classification methods [3,6,14,16]. The discriminative DR methods such as LDA and the linear representation based methods such as SRC usually require that the number of training samples per class cannot be too small, and their performance can be much reduced if the training sample is insufficient. By exploiting more effectively the discriminative information of training sample via learning the projection and dictionary simultaneously, the proposed JDDRDL shows more robust FR capability when the training sample size per class is small, for example 2∼5 samples per class.

The rest of the paper is organized as follows. Section 1 briefly reviews the related work. Section 2 presents in details the JDDRDL algorithm. Section 3 presents the experimental results; and Section 4 concludes the paper.

## 1. Related work

### 1.1. PCA and LDA

As the most representative unsupervised DR method, PCA extracts the eigenvector of the high dimension data, and projects the high dimension data into a linear subspace spanned by leading eigenvectors, seeking a subspace with the maximized variance. PCA is very simple and efficient in reducing the sensitivity to Gaussian noise and some trivial information; however, PCA aims to preserve the global energy of face images but not the discrimination of face images. In contrast, as the most representative supervised DR method, LDA seeks directions which are best for discrimination. LDA finds projections that can minimize the variation of samples in the same class while maximizing the variation between different classes. LDA is effective for classification; however, it is sensitive to the number of training samples per class. In addition, the reduced dimensionality cannot be greater than the number of classes, which limits LDA's applications in practice.

### 1.2. SRC [6] and CRC (collaborative representation classification [26])

The SRC scheme proposed by Wright et al. [6] uses sparse representation for FR. Suppose $\boldsymbol{A}_k=[\boldsymbol{s}_{k,1}, \boldsymbol{s}_{k,2}, \ldots, \boldsymbol{s}_{k,n}]\in\Re^{m\times n}$ is the training dataset of the $k^{\text{th}}$ class, where $\boldsymbol{s}_{k,j}$, $j=1, 2,\ldots, n$, is an $m$-dimensional vector stretched by the $j^{\text{th}}$ sample of class $k$. For a test sample $\boldsymbol{y}\in\Re^m$ from class $k$, generally it can be well approximated as the linear combination of the samples from $\boldsymbol{A}_i$, i.e., $\mathbf{y}\approx\sum_{j=1}^{n}\alpha_{k,j}\mathbf{s}_{k,j}=\mathbf{A}_k\boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k=[\alpha_{k,1},\alpha_{k,2},\ldots,\alpha_{k,n}]^T\in\Re^n$ is the coding vector. Suppose we have $K$ classes, and let $\boldsymbol{A}=[\boldsymbol{A}_1, \boldsymbol{A}_2,\ldots, \boldsymbol{A}_K]$, then the linear representation of $\boldsymbol{y}$ can be written in terms of all training samples as $\boldsymbol{y}\approx\boldsymbol{A\alpha}$, where $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1;\ldots,\boldsymbol{\alpha}_k;\ldots;\boldsymbol{\alpha}_K]=[0,\ldots,0,\alpha_{i,1},\alpha_{i,2},\ldots,\alpha_{i,n},0,\ldots,0]^T$. Clearly, the non-zero element in the coefficient vector could well encode the identity of the test image $\boldsymbol{y}$. In SRC [6], the $l_1$-minimization is used to solve the coding vector: $\hat{\boldsymbol{\alpha}}=\operatorname{argmin}_{\boldsymbol{\alpha}}\{\|\boldsymbol{y}-\boldsymbol{A\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1\}$, where $\lambda$ is a scalar constant. Then classification is made by $identity(\boldsymbol{y})=\operatorname{argmin}_k\{e_k\}$, where $e_k=\|\boldsymbol{y}-\boldsymbol{A}_k\hat{\boldsymbol{\alpha}}_k\|_2$.

The SRC achieves interesting FR results; however, the use of $l_1$-minimization makes it computationally expensive. SRC and its many variants [5,14,16] emphasize the role of $l_1$-norm sparsity in the success of SRC. Very recently, Zhang et al. [26] pointed out that the success of SRC mainly comes from the collaborative representation of the query image by using all the training samples, but not the $l_1$-norm sparsity imposed on the coding vector. Based on this finding, Zhang et al. proposed the collaborative representation based classification (CRC), where the $l_2$-norm is used to regularize the coding coefficients: $\hat{\boldsymbol{\alpha}}=\operatorname{argmin}_{\boldsymbol{\alpha}}\{\|\boldsymbol{y}-\boldsymbol{A\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_2^2\}$. It is shown that when the facial feature dimension is not much less than the number of training samples, CRC could achieve similar FR rates to SRC but the time complexity is enormously reduced.

### 1.3. DR and DL under the SRC framework

It is claimed in [6] that SRC is insensitive to feature extraction when the dimensionality is high enough; however, a well learned DR matrix can lead to a more accurate and stable recognition result. In [14], an orthogonal DR matrix $\boldsymbol{P}$ was learnt under the framework of sparse representation, and it achieves better performance than Eigenfaces and Randomfaces in the SRC scheme. Specifically, the matrix $\boldsymbol{P}$ is learnt via the following objective function based on Leave-One-Out scheme:

$$J_{\mathbf{P},\{\boldsymbol{\beta}_i\}}=\operatorname{argmin}\{\textstyle\sum_{i=1}^{N}(\|\boldsymbol{P}\mathbf{z}_i-\boldsymbol{P}\boldsymbol{A}_i\boldsymbol{\beta}_i\|_F^2+\lambda_1\|\boldsymbol{\beta}_i\|_1)$$
$$+\lambda_2\|\mathbf{A}-\boldsymbol{P}^T\boldsymbol{P}\mathbf{A}\|_F^2\}\quad s.t.\ \boldsymbol{P}\boldsymbol{P}^T=\boldsymbol{I}$$

where $N$ is the number of training samples, $\boldsymbol{z}_i$ is the $i^{\text{th}}$ sample of the training set $\boldsymbol{A}$ and $\boldsymbol{A}_i$ is the set of training samples in $\boldsymbol{A}$ excluding $\boldsymbol{z}_i$. As can be seen from the above objective function, the projection matrix $\boldsymbol{P}$ preserves the energy of training set $\boldsymbol{A}$ while keeping the coding vector of each sample $\boldsymbol{z}_i$ sparse.

In SRC, the original training samples are used as the dictionary to represent the query sample. Intuitively, a more accurate and discriminative representation can be obtained if we could optimize a dictionary from the original training samples. In [16], Yang et al. proposed a "metaface" learning method, where a dictionary $\boldsymbol{D}_k=[\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p]$ of metafaces is learned from each class of training samples $\boldsymbol{A}_k$ under the sparse representation model via optimizing $J_{\boldsymbol{D}_k,\boldsymbol{\Lambda}}=\operatorname{argmin}_{\boldsymbol{D}_k,\boldsymbol{\Lambda}}\|A_k-\boldsymbol{D}_k\boldsymbol{\Lambda}\|_F^2+\lambda\|\boldsymbol{\Lambda}\|_1 s.t.\ \boldsymbol{d}_j^T\boldsymbol{d}_j=1,\quad j=1,\ldots,p$. The metaface dictionary $\boldsymbol{D}_k$ and the associated coefficient matrix $\boldsymbol{\Lambda}$