Linear Subspace Learning based on a Learned Discriminative Dictionary for Sparse Coding

Shibo Gao[1], Yizhou Yu[2] and Yongmei Cheng[1]
[1] *College of Automation, Northwestern Polytechnical University, Xi'an, China*
[2]*Department of Computer Science, The University of Hong Kong, Hong Kong, Hong Kong*

Keywords: Face Recognition, Linear Subspace Learning, Discriminative Dictionary Learning.

Abstract: Learning linear subspaces for high-dimensional data is an important task in pattern recognition. A modern approach for linear subspace learning decomposes every training image into a more discriminative part (MDP) and a less discriminative part (LDP) via sparse coding before learning the projection matrix. In this paper, we present a new linear subspace learning algorithm through discriminative dictionary learning. Our main contribution is a new objective function and its associated algorithm for learning an overcomplete discriminative dictionary from a set of labeled training examples. We use a Fisher ratio defined over sparse coding coefficients as the objective function. Atoms from the optimized dictionary are used for subsequent image decomposition. We obtain local MDPs and LDPs by dividing images into rectangular blocks, followed by blockwise feature grouping and image decomposition. We learn a global linear projection with higher classification accuracy through the local MDPs and LDPs. Experimental results on benchmark face image databases demonstrate the effectiveness of our method.

1 INTRODUCTION

Linear subspace learning (LSL) is a popular method for dimensionality reduction and feature extraction for many pattern recognition tasks, including face recognition (Cai et al., 2007, Huang et al., 2011, Lu et al., 2010). It typically learns an optimal subspace or linear projection according to a task-driven criterion. High-dimensional data can be linearly reduced to lower-dimensional subspaces through LSL. Usually, recognition performance can be improved in such lower-dimensional subspaces (Cai et al., 2007).

There are both unsupervised and supervised LSL methods according to whether they exploit the class label information of train data. Representative techniques of the two classes of LSL methods are Eigenface (Turk et al., 1991) and Fisherface (Belhumeur et al., 1997), respectively. The corresponding algorithms are Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). PCA seeks an optimal subspace with maximal variances while LDA looks for a linear combination of features which characterize or separate two or more classes of objects. A supervised LSL method that utilizes the label information can usually obtain a better discriminative subspace for classification problems. Many supervised LSL methods have been proposed as variants of LDA, including regularized LDA (RLDA) (Lu et al., 2005), and perturbation LDA (PLDA) (Zheng et al., 2009). As shown in (Zheng et al., 2009), both of LDA and RLDA share an assumption that the class empirical mean is equal to its expectation. This assumption may not be valid in practice and a new algorithm, called perturbation LDA (PLDA), is developed. In the algorithm, perturbation random vectors are introduced to learn the effect of the difference between the class empirical mean and its expectation under the Fisher criterion. When the number of training samples is less than the dimensionality of a sample, the within-class scatter matrix in LDA becomes singular, and PCA is often used to reduce the dimensionality before LDA. In (Zheng et al., 2005) a GA-Fisher method is proposed to select the eigenvectors automatically in PCA before LDA. In (Qiao et al., 2009) a sparse LDA is presented to overcome the small sample problem. In (Ji et al., 2008), the authors present a unified framework for generalized LDA via a transfer function.

A common goal shared among existing LSL

Gao S., Yu Y. and Cheng Y..
Linear Subspace Learning based on a Learned Discriminative Dictionary for Sparse Coding.
DOI: 10.5220/0004207905300538
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 530-538
ISBN: 978-989-8565-47-1

Copyright © 2013 SCITEPRESS (Science and Technology Publications, Lda.)

methods is that the learned subspace should be as discriminative as possible. However, every image is a superposition of both discriminative and non-discriminative information. Most existing LSL methods estimate the scatter matrices directly from the original training samples or images. The non-discriminative information in such training samples, such as noise and trivial structures, may interfere with discriminative subspace learning. Different from most existing LSL methods, such as PCA, FLDA/RLDA, LPP and SPP, where the subspace is learned for image decomposition, a new LSL framework is presented in (Zhang et al., 2011) to perform image decomposition for subspace learning. Dictionary learning and sparse coding are used for adaptive image decomposition during the learning stage, where the image is decomposed and the image components are used for guiding subspace learning.

With the development of l_0- and l_1-minimization techniques, sparse coding and dictionary learning have received much attention recently. The dictionary learned in (Zhang et al., 2011) is a generative or reconstructive dictionary which only minimizes reconstruction errors. The atoms of the dictionary do not necessarily have sufficient power to discriminate among data with different class labels. Thus, selecting the most discriminative atoms from such a dictionary as in (Zhang et al., 2011) may not achieve the full potential of sparse coding. On the other hand, several methods have been developed to represent discriminative information during dictionary learning. A discriminative term based on LDA is integrated into the classical reconstructive energy formulation of sparse coding in (Huang et al., 2007, Rodriguez et al., 2008). However, a predefined dictionary instead of a learned dictionary is used in (Huang et al., 2007). In (Yang et al., 2011), a discriminative dictionary is learned based on an objective function combining a variant of Fisher criterion and a reconstruction error. A potential problem with such an approach is that the reconstruction error term may interfere with the Fisher criterion and reduce its power when learning a discriminative dictionary. A formulation of the classification error of a linear SVM has also been incorporated into dictionary learning (Jiang et al., 2011, Zhang et al., 2010). Other efforts along this direction include multi-class dictionary optimization for gaining discriminative power in texture analysis (Mairal et al., 2008). A compact dictionary is learned from affine-transformed input images to increase discriminative information (Kulkarni et al., 2011). In (Mairal et al., 2012) a task-driven supervised dictionary learning method is proposed where dictionary learning relies on a subgradient method to perform a nonconvex optimization. Note that learning discriminative dictionaries based on the SVM error term is not well suited for problems that involve image patches because even images from different classes may share similar patches, whose class labels would be very hard to determine.

In this paper, we present a new linear subspace learning method based on sparse coding using a novel technique for discriminative dictionary learning. We use a Fisher ratio defined over sparse coding coefficients as the objective function for optimizing a discriminative dictionary while the condition that the sparse coding coefficients should minimize the reconstruction error is imposed as a constraint. Atoms that can achieve a higher Fisher ratio of the sparse coding coefficients are considered better. Therefore, discriminative information is emphasized during the atom construction process. In our decomposed image, the more discriminative part (MDP) has a larger fisher ratio than the one obtained from a reconstructive dictionary. We further obtain local MDPs and LDPs by dividing images into rectangular blocks, followed by blockwise feature grouping and image decomposition. A more effective global MDP and LDP can be obtained by concatenating these local MDPs and LDPs. We learn a global linear projection with higher classification accuracy through the global MDPs and LDPs. Experimental results on benchmark face image databases demonstrate the effectiveness of our method. The flowchart of the proposed LSL method is shown in Fig. 1.

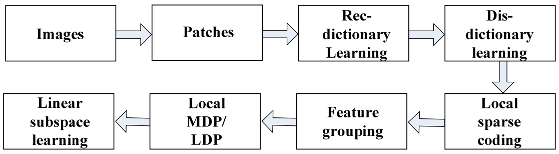


Figure 1: The flowchart of the proposed method.

2 IMAGE DECOMPOSITION AND RECONSTRUCTION VIA SPARSE CODING

The sparse representation model is a modern method for image decomposition and reconstruction, which have been used in many image-related applications, such as image restoration and feature selection. The sparse representation of a signal over an over-complete dictionary is achieved by optimizing an objective function that includes two terms: one

SCIENCE AND TECHNOLOGY PUBLICATIONS

measures the signal reconstruction error and the other measures the coefficients' sparsity (Huang et al., 2007). Suppose that the data $\boldsymbol{X} = [x_1, x_2, \dots, x_n] \in \mathrm{R}^{m \times n}$ (n is the number of samples, m is the number of dimensions) admits a sparse approximation over an over-complete dictionary $\boldsymbol{D} = [d_1, d_2, \dots, d_K] \in \mathrm{R}^{m \times K}$ (with $K>m$) with K atoms. Then X can be approximately represented as a linear combination of a sparse subset of atoms from $\boldsymbol{D}$. The over-complete dictionary $\boldsymbol{D}$ can be obtained by solving the following optimization problem: $\arg\min_{\boldsymbol{D}, \{\alpha_i\}_{i=1}^{i=n}} \sum_{i=1}^{n} \|x_i - \boldsymbol{D}\alpha_i\|_F^2$, $s.t. \|\alpha_i\|_0 \le \varepsilon$, where α_i is a sparse coefficient vector, and each column of $\boldsymbol{D}$ is an atom represented as a unit vector. In practice, the l_1 norm replaces the l_0 norm during optimization, that is,

$$\arg\min_{\boldsymbol{D}, \{\alpha_i\}_{i=1}^{i=n}} \left\{ \sum_{i=1}^{n} \|x_i - \boldsymbol{D}\alpha_i\|_F^2 + \lambda \|\alpha_i\|_1 \right\} \tag{1}$$

The weight λ controls the trade-off between the reconstruction accuracy and the sparseness of the coefficient vectors. The cost function given above is non-convex with respect to both $\boldsymbol{D}$ and α_i . However, it is convex when one of them is fixed. Thus, this problem can be approached by alternating between learning $\boldsymbol{D}$ while fixing α_i and inferring α_i while fixing $\boldsymbol{D}$. We can see that a data sample x can be decomposed into $x \approx \boldsymbol{D}\alpha = \sum_{k=1}^{K} \alpha(k) d_k$. If the sparse coefficient vector α for the data item x is known under the dictionary $\boldsymbol{D}$, then the pre-image can be reconstructed by $\hat{x} = \boldsymbol{D}\alpha$. The sparse coefficient vector α and the dictionary $\boldsymbol{D}$ contain the most information about the data item x, therefore, they have been used for tasks such as recognition, feature extraction and image restoration.

3 DISCRIMINATIVE DICTIONARY LEARNING

In this section, we introduce a new objective function and its associated algorithm for learning an overcomplete discriminative dictionary from a set of labeled training image patches. Supervised dictionary learning for achieving a goal other than data reconstruction has proven to be a hard problem, especially when the objective function for dictionary learning is not differentiable everywhere in l_1 norm restraint. Our algorithm relies on the subgradient method to obtain optimized atoms. One of the key contributions of our work is to train a discriminative dictionary for subsequent feature grouping.

Note that Fisher's criterion is motivated by the intuitive idea that data samples from multiple classes are maximally separated when samples from different classes are distributed as far away from each other as possible and samples from the same class are distributed as close to each other as possible. Because of the reconstructive capability of an over-complete dictionary, the degree of separation among data samples can be indicated by the degree of separation among the sparse coefficient vectors coding the data samples using the dictionary. Thus, discriminative components of data samples can be discovered by seeking a dictionary whose corresponding sparse coefficient vectors achieve a high Fisher's ratio. Therefore, we use the Fisher's ratio defined over the sparse coefficient vectors as the objective function for learning a discriminative dictionary.

Let $x_i \in \mathrm{R}^m$ ($i = 1, 2, \dots, n_{C_k}$) be a training sample, an m-dimensional vector formed by image patches of size $\sqrt{m} \times \sqrt{m}$ from images in the C_k -th class, n_{C_k} is the number of training samples in the C_k -th class. The Fisher's criterion is defined as $Fisher(x) = tr(\boldsymbol{S}_B(x)) / tr(\boldsymbol{S}_W(x))$, where $\boldsymbol{S}_W(x)$ is the within-class scatter matrix, $S_B(x)$ is the between-class scatter matrix, which are defined as $\boldsymbol{S}_W(x) = \sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left((x_i - \mathrm{m}_k)(x_i - \mathrm{m}_k)^T \right)_{x_i \in C_k}$ and $\boldsymbol{S}_B(x) = \sum_{k \in C} n_{C_k} (\mathrm{m}_k - \mathrm{m})(\mathrm{m}_k - \mathrm{m})^T$, where m_k is the mean of training samples in class C_k , and m is the mean of all training samples. We can see that

$$Fisher(x) = tr(\boldsymbol{S}_B(x)) / tr(\boldsymbol{S}_W(x)) = \sum_{k \in C} n_{C_k} \|\mathrm{m}_k - \mathrm{m}\|_F^2 \Big/ \sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\|(x_i - \mathrm{m}_k)\|_F^2 \right)_{x_i \in C_k} .$$

Based on the Fisher's criterion, we define discriminative dictionary learning as follows:

$$\arg\min_{\boldsymbol{D}} J\left(\alpha^*(x, \boldsymbol{D})\right) = \frac{\sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\|\alpha_i^* - \mu_k\|_F^2 \right)_{\alpha_i^* \in C_k}}{\sum_{k \in C} n_{C_k} \|\mu_k - \mu\|_F^2} \tag{2}$$

s.t. $\alpha_i^*(x, \boldsymbol{D}) = \arg\min_{\alpha} \|x_i - \boldsymbol{D}\alpha\|_F^2 + \lambda \|\alpha\|_1$

where $\alpha_i^*(x, \boldsymbol{D})$ represents the optimal sparse coefficient vector coding the training sample x using

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

the learned dictionary $\boldsymbol{D}$, $\mu_k = \left(\frac{1}{n_{C_k}} \sum_{i=1}^{n_{C_k}} \alpha_i^* \right)_{i \in C_k}$ is the mean sparse coefficient vector for training samples in class C_k , $\mu = \frac{1}{n} \sum_{k=1}^{n} \alpha_k^*$ is the mean sparse coefficient vector, $n = \sum_{k \in C} n_{C_k}$ is the total number of training samples. Obviously, the numerator of the above objective function, $\sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\left\| \alpha_i^* - \mu_k \right\|_F^2 \right)_{\alpha_i^* \in C_k}$, represents the intra-class compactness, and the denominator, $\sum_{k \in C} n_{C_k} \left\| \mu_k - \mu \right\|_F^2$, represents the inter-class separability. Note that $Fisher\left(\alpha^*\right) = 1/J\left(\alpha^*(\mathrm{x}, \boldsymbol{D})\right)$ is the Fisher's ratio defined over the optimal sparse coefficient vectors. So $\min J\left(\alpha^*(\mathrm{x}, \boldsymbol{D})\right) \Rightarrow \max Fisher\left(\alpha^*\right)$. Note that in the optimization defined in equation (2), the discriminative Fisher ratio and the reconstruction term are separated, one as the objective function and the other as the constraint. Such a scheme guarantees that there is least interference between the two and the power of the Fisher criterion is fully realized.

Our method learns an entire discriminative dictionary from labeled training data using the Fisher criterion. The optimal sparse coefficient vector of a training sample is actually a function of dictionary $\boldsymbol{D}$ due to $\alpha_i^*(x, \boldsymbol{D}) = \arg\min_{\alpha} \left\| x_i - \boldsymbol{D}\alpha \right\|_F^2 + \lambda \left\| \alpha \right\|_1$. Therefore, dictionary $\boldsymbol{D}$ is the only variable of the objective function $J\left(\alpha^*(x, \boldsymbol{D})\right)$. We will use $F(\boldsymbol{D})$ as a short notation in place of $J\left(\alpha^*(x, \boldsymbol{D})\right)$ when necessary. In (Boureau et al., 2010, Mairal et al., 2012, Yang et al., 2010), the subgradient method was used to optimize functions of $\alpha_i^*(x, \boldsymbol{D})$ with respect to D (see the **APPENDIX**). In our problem, the subgradient of $F(\boldsymbol{D})$ with respect to $\boldsymbol{D}$ can in turn be computed using the chain rule

$$
\begin{aligned}
\frac{\partial F(\boldsymbol{D})}{\partial \boldsymbol{D}} &= \sum_i \frac{\partial J\left(\alpha^*\right)}{\partial \alpha_i^*} \frac{\partial \alpha_i^*}{\partial \boldsymbol{D}} \\
&= \sum_i \left(-\boldsymbol{D} \beta_i \alpha_i^{*T} + \left(x_i - \boldsymbol{D} \alpha_i^* \right) \beta_i^T \right) \qquad (3) \\
&= -\boldsymbol{D}\boldsymbol{B}\boldsymbol{A}^{*T} + \left(\boldsymbol{X} - \boldsymbol{D}\boldsymbol{A}^* \right) \boldsymbol{B}^T
\end{aligned}
$$

where $\boldsymbol{A}^* = \left[\alpha_1^*, \alpha_2^*, \ldots \alpha_{C_k}^*, \alpha_{1+C_k}^*, \ldots, \alpha_{\sum_k C_k}^* \right] \in \mathrm{R}^{K \times n}$, $\boldsymbol{B} = \left[\beta_1, \beta_2, \ldots \beta_{C_k}, \beta_{1+C_k}, \ldots, \beta_{\sum_k C_k} \right] \in \mathrm{R}^{K \times n}$, β_i is composed from $(\beta_i)_{i,\Lambda}$ and $(\beta_i)_{i,\Lambda^C}$. $(\beta_i)_{i,\Lambda} = \left(\boldsymbol{D}_{i,\Lambda}^T \boldsymbol{D}_{i,\Lambda} \right)^{-1} \left(\frac{\partial J(\alpha_i^*)}{\partial \alpha_i^*} \right)_{i,\Lambda}$, $(\beta_i)_{i,\Lambda^C} = 0$, where $(\cdot)_{i,\Lambda} = \left\{ j \in \{1, \ldots, K\} : \alpha_i^*[j] \neq 0 \right\}$ is the active set, denoting the indices of the nonzero coefficients within the sparse vector $\alpha_i^*(x, \boldsymbol{D})$, and $(\cdot)_{i,\Lambda^C} = \left\{ j \in \{1, \ldots, K\} : \alpha_i^*[j] = 0 \right\}$ is the non-active set. Since

$$
\begin{aligned}
\frac{\partial J\left(\alpha_i\right)}{\partial \alpha_i} &= \frac{\partial \left(\frac{\sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\left\| \alpha_i - \mu_k \right\|_F^2 \right)_{\alpha_i \in C_k}}{\sum_{k \in C} n_{C_k} \left\| \mu_k - \mu \right\|_F^2} \right)}{\partial \alpha_i} \qquad (4) \\
&= \frac{2\left(\alpha_i - \mu_k\right)}{\sum_{k \in C} n_{C_k} \left\| \mu_k - \mu \right\|_F^2} - \frac{\sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\left\| \alpha_i - \mu_k \right\|_F^2 \right)_{\alpha_i \in C_k}}{\left(\sum_{k \in C} n_{C_k} \left\| \mu_k - \mu \right\|_F^2 \right)^2} \left(2\left(\mu_k - \mu \right) \right)
\end{aligned}
$$

We have

$$
(\beta_i)_{i,\Lambda} = \left(\boldsymbol{D}_{i,\Lambda}^T \boldsymbol{D}_{i,\Lambda} \right)^{-1}
\left(\begin{array}{l}
\frac{2\left(\alpha_i^* - \mu_k\right)_{i,\Lambda}}{\sum_{k \in C} n_{C_k} \left\| \left(\mu_k - \mu \right)_{i,\Lambda} \right\|_F^2} \\
- \frac{\sum_{k \in C} \sum_{i=1}^{n_{C_k}} \left(\left\| \left(\alpha_i^* - \mu_k \right)_{i,\Lambda} \right\|_F^2 \right)_{\alpha_i^* \in C_k}}{\left(\sum_{k \in C} n_{C_k} \left\| \left(\mu_k - \mu \right)_{i,\Lambda} \right\|_F^2 \right)^2} \left(2\left(\mu_k - \mu \right)_{i,\Lambda} \right)
\end{array} \right) \qquad (5)
$$

In summary, the subgradient of $F(\boldsymbol{D})$ at $\boldsymbol{D}$ can be computed as

$$
\boldsymbol{G} = -\boldsymbol{D}\boldsymbol{B}\boldsymbol{A}^{*T} + \left(\boldsymbol{X} - \boldsymbol{D}\boldsymbol{A}^* \right) \boldsymbol{B}^T \qquad (6)
$$

Since the subgradient method can obtain a search direction at a certain point during optimization, it can be used in convex, quasiconvex and nonconvex problems (Burachik et al., 2010, Neto et al., 2011, Yang et al., 2010). Here the dictionary $\boldsymbol{D}$ can be updated by the subgradient $\boldsymbol{G}$. That is

$$
\boldsymbol{D}^{(t)} = \boldsymbol{D}^{(t-1)} - \rho_t \frac{\boldsymbol{G}^{(t-1)}}{\left\| \boldsymbol{G}^{(t-1)} \right\|_F} \qquad (7)
$$

where ρ_t is a learning rate, t is the current iteration step. The step size of the learning rate should be chosen carefully. In this paper, the learning rate ρ_t is calculated from $\rho_t = \rho_0 / \left(1 + t/T\right)$, where ρ_0 is the initial learning rate, ρ_t is the current learning rate,

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

T is a predefined parameter, and the initial dictionary is learned using equation (1). Figure 2 shows the Fisher ratio versus the number of iterations on image patches from the CMU PIE database. We can see that the subgradient-based optimization is a process with oscillatory convergence.

Once the dictionary $\boldsymbol{D}$ has been learned, a data sample can be decomposed into components using the atoms in $\boldsymbol{D}$ as follows, $x_j \approx \boldsymbol{D}\alpha_j = x_{j,1} + x_{j,2} + \ldots + x_{j,K}$, where $x_{j,k} = \alpha_j(k) d_k$

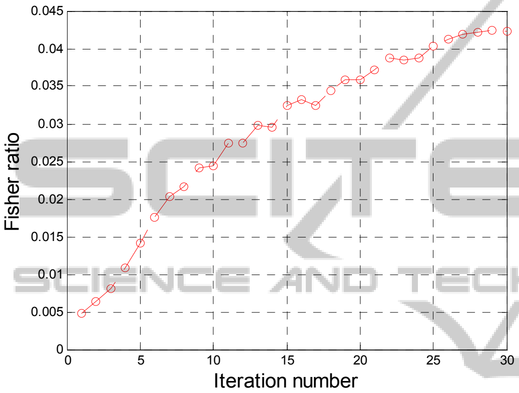


Figure 2: Fisher ratio versus the number of iterations on a subset of the CMU PIE database.

4 LSL VIA DISCRIMINATIVE DICTIONARY LEARNING AND BLOCKWISE DECOMPOSITION

4.1 Patch-based Dictionary Learning

The relatively high dimensionality of an image as well as the usual relatively small number of training images prevent us from learning a redundant dictionary directly from a set of training images under the sparse coding framework. Same to (Zhang et al., 2011), a patch-based scheme is used to learn a dictionary. As mentioned above, the defined objective function (3) is very fit for patch-based discriminative dictionary learning. Each training image I_i is partitioned into overlapping patches. The complete set of patches from all training images is denoted as $\boldsymbol{X}=[x_1, x_2, \ldots, x_n]$, in which *h* is the total number of patches. A discriminative dictionary is learned from *T* following the optimization framework presented in the previous section. Each patch t_j can be reconstructed through the atoms in the learned dictionary as follows, $x_j \approx \boldsymbol{D}\alpha_j = \alpha_j(1)d_1 + \alpha_j(2)d_2 + \ldots + \alpha_j(K)d_K$.

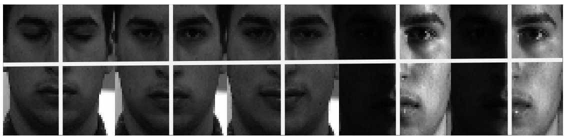


Figure 3: Partition of training images into blocks.

4.2 Blockwise Feature Grouping

We also partition an image I_i into relatively large blocks and perform blockwise feature grouping. Note that each block includes multiple image patches. If an image is divided into *L* blocks, for each patch x_j^l in block *l*, $x_j^l \approx D\alpha_j^l = \sum_{k=1}^{K} \alpha_j^l(k) d_k = \sum_{k=1}^{K} x_{j,k}^l$ (*l=1,2,...,L*). By placing together all patch components $x_{j,k}^l$, corresponding to every atom d_k across the entire block *l*. This image block can be by combining these patches: $R_i^l \approx R_{i,1}^l + R_{i,2}^l + \ldots + R_{i,K}^l$. Each of these components is treated as a feature, and all *K* features are separated into two groups, a more discriminative part (MDP) and a less discriminative part (LDP), according to the magnitude of the following Fisher ratio,

$$f_z^l = \frac{\sum_{k=1}^{C} n_{C_k} \left\| \overline{R}_z^l - \overline{R}_{z,k}^l \right\|_F^2}{\sum_{k=1}^{C} \sum_{i=1}^{n_{C_k}} \left(\left\| R_{i,z}^l - \overline{R}_{z,k}^l \right\|_F^2 \right)_{R_{i,z}^l \in C_k}}, \ l=1,..,L,\ z=1,..K \quad (8)$$

where $\overline{R}_z^l$ the mean of the *z*-th feature of block *l* from all training images, and by $\overline{R}_{z,k}^l$ the mean of the *z*-th feature of block *l* from images that belong to class *k*.

Those features of block *l* that have larger f_z^l are added together to form the MDP image of block *l*, the LDP image of block *l* is formed by original images with the MDP image subtracted. In this way, we compute the MDP and LDP of every block within every training image. We denote by $R_i^{l,a}$ the MDP of block *l* within image I_i, and $R_i^{l,b}$ the LDP of block *l* within image I_i.

Finally, we define the MDP and LDP of a training image by concatenating the MDPs and LDPs of all blocks within the image. That is, let $I_i^{L,a}$ and $I_i^{L,b}$ be the MDP and LDP of image I_i when it is partitioned into *L* blocks, then $I_i = I_i^{L,a} + I_i^{L,b}$, where $I_i^{L,a} = [R_i^{1,a}, R_i^{2,a}, \ldots, R_i^{L,a}]$ and $I_i^{L,b} = [R_i^{1,b}, R_i^{2,b}, \ldots, R_i^{L,b}]$.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

The superscript L in $I_i^{L,a}$ and $I_i^{L,b}$ is due to the fact that the MDP and LDP of an image are dependent on the number of blocks in the image. When the number of blocks changes, the MDP and LDP may also change. Here we assume when the number of blocks is fixed, only one particular way is allowed to partition an image into blocks.

4.3 Subspace Learning

Once the MDP and LDP of an image have been defined, the whole dataset Q can be written as $Q = Q^a + Q^b$, where $Q^a = \left[I_1^{L,a}, I_2^{L,a}, \ldots, I_n^{L,a}\right]$ and $Q^b = \left[I_1^{L,b}, I_2^{L,b}, \ldots, I_n^{L,b}\right]$. As suggested by [11], with a learned linear projection matrix P, that is $\boldsymbol{PQ} = \boldsymbol{PQ}^a + \boldsymbol{PQ}^b$, the features in Q^a should be preserved and the features in Q^b should be suppressed after the projection by $\boldsymbol{P}$. Thus, the optimal linear projection should be

$$\arg\max_{\boldsymbol{P}} \frac{\mathrm{tr}\left\{\boldsymbol{P}\boldsymbol{S}_B^a\boldsymbol{P}\right\}}{\gamma\,\mathrm{tr}\left\{\boldsymbol{P}\boldsymbol{S}_W^a\boldsymbol{P}\right\} + (1-\gamma)\,\mathrm{tr}\left\{\boldsymbol{P}\boldsymbol{S}^b\boldsymbol{P}\right\}} \qquad (9)$$

where $\boldsymbol{S}_B^a$ is the MDP between-class scatter matrix, $\boldsymbol{S}_W^a$ is the MDP within-class scatter matrix, and $\boldsymbol{S}^b$ is the scatter matrix of the less discriminative parts in all images. The exact definitions of these matrices can be found in (Zhang et al., 2011).

Remarks. In comparison with the linear subspace learning method in (Zhang et al., 2011), our algorithm exhibits a few advantages. First, different from the reconstructive dictionary used in (Zhang et al., 2011), we train a discriminative dictionary. Atoms in a discriminative dictionary can better expose discriminative components in an image. Second, the objective function we use for training the dictionary is a Fisher ratio, which is perfectly consistent with the subsequent feature grouping criterion, which is also based on a Fisher ratio. In contrast, the dictionary training criterion in (Zhang et al., 2011) is not consistent with the feature grouping criterion. Third, instead of selecting the same features for an entire image, we perform blockwise feature grouping. Our selected features are optimally discriminative for each local image block and may vary from block to block. Such a spatially varying strategy can potentially discover more discriminative image components.

Table 1: Outline of our algorithm.

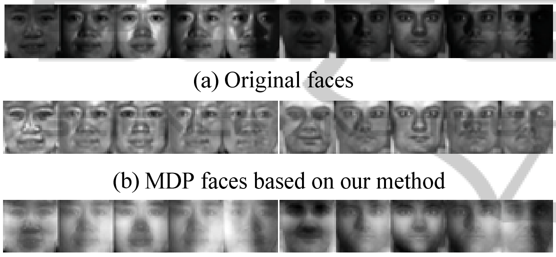
```
Linear Subspace Learning via Discriminative Dictionary

Initialize: face images, t=1, learning rate $\{\rho_t\}_t^T$,
            number of iterations $t_{\max}$.
Output: linear projection P.
Initialize $\boldsymbol{D}^1$ according to equation (1);
Repeat
    Compute coefficients $\alpha_i^t$ via
    $\arg\min_{\alpha_i} \|x_i - \boldsymbol{D}^t\alpha_i\|_F^2 + \lambda\|\alpha_i\|_1$;
    Compute Fisher ratio $Fisher\left(\alpha^{*(t)}\right)$;
    if $Fisher\left(\alpha^{*(t)}\right) = \max\left\{Fisher\left(\alpha^{*(t)}\right)\right\}_{t\in T_0}$,
        $\boldsymbol{D}^0 = \boldsymbol{D}^{(t+1)}$;
    Compute $\boldsymbol{G}^t$ according to equation (6);
    Compute $\beta_i^t$ within $\boldsymbol{B}^t$ according to equation (5);
        Update $\boldsymbol{G}^t = \frac{\boldsymbol{G}^t}{\|\boldsymbol{G}^t\|_F}$;
        Update $\boldsymbol{D}^{(t+1)} = \boldsymbol{D}^{(t)} - \rho_t\boldsymbol{G}^t$;
        Update $\boldsymbol{D}^{t+1} = \frac{\boldsymbol{D}^{t+1}}{\|\boldsymbol{D}^{t+1}\|_F}$;
        t=t+1;
If $t < t_{\max}$, continue repeat; Else stop repeat;
End
Compute coefficients $\alpha_i$ via
$\arg\min_{\alpha_i} \|x_i - \boldsymbol{D}^0\alpha_i\|_F^2 + \lambda\|\alpha_i\|_1$;
Divide images into blocks;
Feature grouping on each block according to equation (8);
Compute linear projection P according to equation (9).
```

5 EXPERIMENTAL RESULTS

All experiments were carried out on an Intel i7 3.40GHz processor with 16GB RAM. Our LSL method is based on sparse coding using the learned discriminative dictionary (DSSCP). The feature-sign search algorithm (Lee et al., 2006) was used for sparse coding during dictionary learning. The performance of DSSCP has been evaluated on two representative facial image databases: the CMU PIE database, and the ORL database. Three state-of-the-art existing methods, LDA (Belhumeur et al., 1997),

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

PLDA (Zheng et al., 2009) and SSCP (Zhang et al., 2011), are used for comparison. Face images with 32×32 pixels were used in our experiments. Each face image was flattened and normalized to a unit vector. During dictionary learning, every image was partitioned into8x8 patches. Our dictionary has 64 atoms, each of which has the same size as an image patch. In the subgradient-based optimization, the initial learning rate ρ_0 was set to 0.02, the parameter T was set to 10, and the number of iterations was set to 30.The number of blocks in an image, L, was set to 4. And 80% atoms in the dictionary were used to form the MDP during feature grouping. Classification was performed using the simple nearest neighbor (NN) classifier. The parameter γ was chosen with10-fold cross-validation on the training set based on the recognition rate.



(a) Original faces

(b) MDP faces based on our method

(c) LDP faces based on our method

Figure 4: Training images of two subjects from the CMU PIE database (First five columns are for one subject, last five columns for another subject). The average Fisher ratio of the MDP of training images by the method in (Zhang et al., 2011) is 0.7554, while the corresponding average Fisher ratio by our method is 0.8291.

5.1 CMU PIE Database

CMU PIE face database (Sim et al., 2003) contains 41368 face images of 68 people, each person with 13 different poses, 43 different illumination conditions, and 4 different expressions. In our experiments, we chose images with one near frontal pose (C27) and all different illumination conditions and expressions. There are 49 near frontal images for every subject.

We chose 30 individuals from the 68 people for our experiments. All the face images were preprocessed with histogram equalization and normalization. First, the images were split into two groups. There are 25 images in group 1 for each subject while 24 images in group 2 for each. We randomly chose 5 images per subject from group 1 images for training, and 5 images per subject from group 2 for testing. Figure 4 shows examples of the original training images, the extracted MDP images and LDP images. PCA was used to reduce the number of dimensions of the total covariance matrix when learning the projection $\boldsymbol{P}$ in all the four methods. The reported recognition rate is the average over 50 runs.

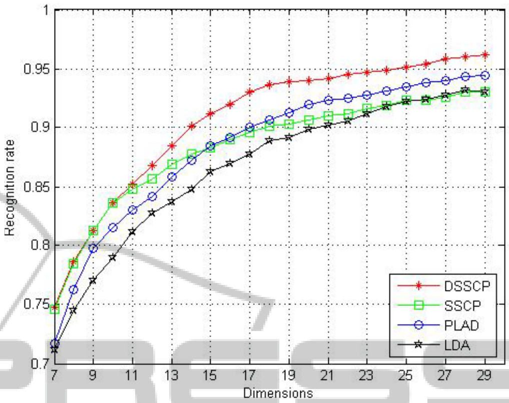


Figure 5: Recognition rates achieved by different methods on the CMU PIE database versus dimensionality of the linear subspace.

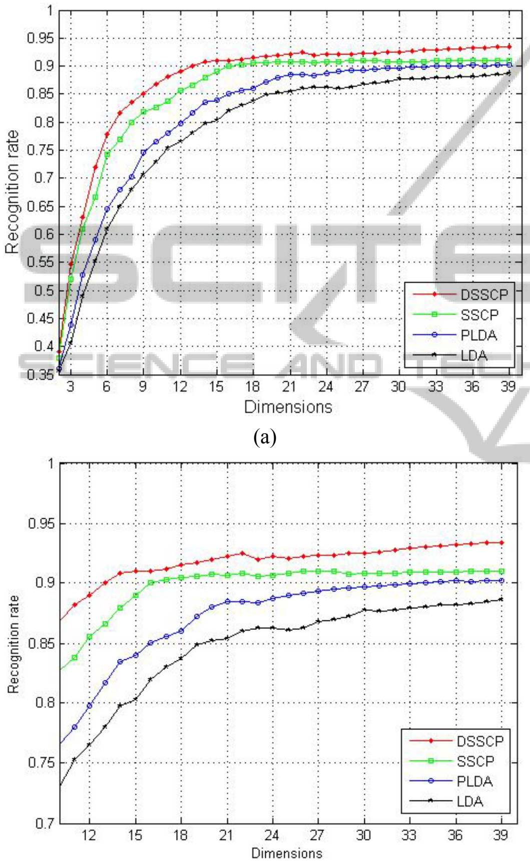
Figure5 summarizes recognition performance of various algorithms used in our experiments. From this figure, we can see that our DSSCP algorithm overall performs better than LDA, PLDA and SSCP. The advantage of our algorithm becomes more obvious at larger dimensions. This comparison indicates image decomposition in general is beneficial for linear subspace learning based on LDA. Furthermore, if more discriminative information is extracted during image decomposition as in our algorithm, we can obtain better linear projections during subspace learning.

5.2 Experiments on the ORL Database

ORL face database contains 400 images of 40 individuals. There are ten different images for each subject. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The used ORL database is available from http://www.zjucadcg.cn/dengcai/Data/FaceData.html. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). We selected 3 images per subject for training, and 5 images for testing.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 6: Face images in the ORL database.



(a)

(b)

Figure 7: Recognition rates achieved by different methods on the ORL database vs. dimensionality of the linear subspace. (b) shows an enlarged portion of (a).

The average recognition rates (over 50 runs) versus subspace dimensionality are shown in Fig. 7. Fig.7 summarizes results from the algorithms tested in our experiments. From the above figure, we can see that DSSCP is better than LDA, PLDA and SSCP. That means the decomposition of image is useful for linear subspace learning based on LDA. We can obtain a better linear subspace projection, if the discriminative information is utilized.

6 CONCLUSIONS

In this paper, we have presented a linear subspace learning algorithm through learning a discriminative dictionary. Our main contributions include a new objective function based on a Fisher ratio over sparse coding coefficients, and its associated algorithm for learning an overcomplete discriminative dictionary from a set of labeled training examples. We further obtain local MDPs and LDPs by dividing images into rectangular blocks, followed by blockwise feature grouping and image decomposition. We learn a global linear projection through the local MDPs and LDPs. Experiments and comparisons on benchmark face recognition datasets have demonstrated the effectiveness of our method. Our future work includes exploring both theoretically and empirically the structure of the learned dictionary with respect to different datasets and dictionary size. In addition, we plan to investigate kernel subspace learning via discriminative dictionaries.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

REFERENCES

Belhumeur, P., Hespanha, J. and Kriengman, D. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *PAMI*, 19 (7): 711-720.

Boureau, Y., Bach, F., LeCun, Y., and Ponce J. (2010). Learning Mid-level Features for Recognition. In *CVPR*: 2559-2566.

Burachik, R., Kaya, C., and Mammadov, M. (2010). An Inexact Modified Subgradient Algorithm for Nonconvex Optimization. Computational *Optimization and Applications*, 45(1): 1-24.

Cai, D., He, X., Hu, Y., Han, J., and Huang, T. (2007). Learning a Spatially Smooth Subspace for Face Recognition. In *CVPR*: 1-7.

Huang, D., Storer, M., Torre, F., and Bischof, H. (2011). Supervised Local Subspace Learning for Continuous Head Pose Estimation. In *CVPR*: 2921-2928.

Huang, K., Aviyente, S. (2007). Sparse Representation for Signal Classification. In *NIPS*: 609-616.

Ji, S. Ye, J. (2008). A Unified Framework for Generalized Linear Discriminant Analysis. In *CVPR*: 1-7.

Jiang, Z., Lin, Z. and Davis, L. (2011). Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*: 1697-1704.

Kulkarni, N., Li, B. (2011). Discriminative Affine Sparse Codes for Image Classification. In *CVPR*: 1609-1616.

Lee, H., Battle A., Raina, R., and Ng, A. Y. (2006). Efficient Sparse Coding Algorithms. In *NIPS*: 801-808.

Lu, J., and Tan, Y. (2010). Cost-Sensitive Subspace Learning for Face Recognition. In *CVPR*: 2661-2666.

Lu, J., Plataniotis, K., Venetsanopoulos, A. (2005).

Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition. *Pattern Recognition Letters*, 26(2): 181-191.

Mairal, J., Bach, F., Ponce, J. (2012). Task-Driven Dictionary Learning. *PAMI*, 34(4): 791-804.

Mairal, J., Bach, F., Ponce, J., Sapiro, G. and Zisserman A. (2008). Discriminative Learned Dictionaries for Local Image Analysis. In *CVPR*: 1-8.

Neto, J., Lopes, J., Travaglia, M. (2011). Algorithms for Quasiconvex Minimization. *Optimization*, 60(8-9): 1105-1117.

Qiao, Z., Zhou, L., Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data. *IAENG International Journal of Applied Mathematics*, 39(1): 1-13.

Rodriguez, F., Sapiro, G. (2008). Sparse Representations for Image Classification: Learning Discriminative and Reconstructivenon-Parametric Dictionaries. http://www.ima.umn.edu/preprints/jun2008/2213.pdf.

Sim, T., Baker, S., Bsat, M. (2003). The CMU Pose, Illumination, and Expression Database. *PAMI*, 25(12): 1615-1618.

Turk, M., and Pentland, A. (1991). Eigenfaces for Recognition. *J. Cognitive Neuroscience*, 3(1):71-86.

Yang, J., Yu, K., and Huang, T. (2010). Supervised Translation-Invariant Sparse Coding. In *CVPR*: 3517-3524.

Yang, M., Zhang, L., Feng, X., Zhang, D. (2011). Fisher Discrimination Dictionary Learning for Sparse Representation. In *ICCV*, 2011: 543-550.

Zhang, L., Zhu, P., Hu, Q., Zhang, D. (2011). A Linear Subspace Learning Approach via Sparse Coding. In *ICCV*: 755-761.

Zhang, Q., Li, B. (2010). Discriminative K-SVD for Dictionary Learning in Face Recognition. In *CVPR*: 2691-2698.

Zheng, W., Lai, J., Yuen, P. (2005). GA-fisher: A New LDA-based Face Recognition Algorithm with Selection of Principal Components. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5): 1065-1078.

Zheng, W., Lai, J. H., Yuen, P. C., Li, S. Z. (2009). Perturbation LDA: Learning the Difference between the Class Empirical Mean and Its Expectation. *Patten Recognition*, 42(5): 764-779.

APPENDIX

The differentiation of a sparse coefficient vector α^* with respect to $\boldsymbol{D}$ is shown in (Mairal et al., 2008, Mairal et al., 2012). Since $\alpha^*_\Lambda = \left(\boldsymbol{D}_\Lambda^T \boldsymbol{D}_\Lambda\right)^{-1}\left(\boldsymbol{D}_\Lambda^T x - \lambda s_\Lambda\right)$, where s_Λ in $\{-1;+1\}^{|\Lambda|}$ carries the signs of α^*_Λ , its subgradient can be computed once α^* known. It is $\frac{\partial\left(\alpha^*_\Lambda\right)_k}{\partial\left(\boldsymbol{D}_\Lambda\right)_{ij}} = \left(x - \boldsymbol{D}\alpha^*\right)_i W_{jk} - \left(\alpha^*_\Lambda\right)_j C_{jk}$, where $W = \left(\boldsymbol{D}_\Lambda^T \boldsymbol{D}_\Lambda\right)^{-1}, C = W\boldsymbol{D}_\Lambda^T$, and $\left(\alpha^*_\Lambda\right)_k$ denotes the k-th nonzero component of α^* . The subgradient of an objective function $h(\alpha^*)$ with respect to $\boldsymbol{D}$ is given as follows (Mairal et al., 2012), $\frac{\partial h(\alpha^*)}{\partial \boldsymbol{D}} = \frac{\partial h(\alpha^*)}{\partial \alpha^*}\frac{\partial \alpha^*}{\partial \boldsymbol{D}} = -\boldsymbol{D}\beta\alpha^{*T} + \left(x - \boldsymbol{D}\alpha^*\right)\beta^T$, where $(\beta)_\Lambda = \left(\boldsymbol{D}_\Lambda^T \boldsymbol{D}_\Lambda\right)^{-1}\frac{\partial h(\alpha^*)}{\partial \alpha^*_\Lambda}, (\beta)_{\Lambda^C} = 0$.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS