Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Discriminative Analysis Dictionary Learning

Jun Guo[1*], Yanqing Guo[1], Xiangwei Kong[1], Man Zhang[2], and Ran He[2,3]
[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[2] The Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Beijing 100190, China
guojun@mail.dlut.edu.cn, {guoyq,kongxw}@dlut.edu.cn, {zhangman,rhe}@nlpr.ia.ac.cn

Abstract

Dictionary learning (DL) has been successfully applied to various pattern classification tasks in recent years. However, analysis dictionary learning (ADL), as a major branch of DL, has not yet been fully exploited in classification due to its poor discriminability. This paper presents a novel DL method, namely Discriminative Analysis Dictionary Learning (DADL), to improve the classification performance of ADL. First, a code consistent term is integrated into the basic analysis model to improve discriminability. Second, a triplet-constraint-based local topology preserving loss function is introduced to capture the discriminative geometrical structures embedded in data. Third, correntropy induced metric is employed as a robust measure to better control outliers for classification. Then, half-quadratic minimization and alternate search strategy are used to speed up the optimization process so that there exist closed-form solutions in each alternating minimization stage. Experiments on several commonly used databases show that our proposed method not only significantly improves the discriminative ability of ADL, but also outperforms state-of-the-art synthesis DL methods.

1 Introduction

The success of sparse representation (SR) pushes forward the research of dictionary learning (DL). In SR, a desired dictionary learned from data often outperforms a set of predefined bases in pattern classification tasks. One popular line of research in DL aims to learn a synthesis dictionary with specific promotion functions. Synthesis dictionary learning is widely used but time-consuming. Hence, as its dual model, analysis model has drawn much attention recently (Rubinstein, Bruckstein, and Elad 2010).

Analysis dictionary learning (ADL) aims to learn a transformation instead of utilizing off-the-shelf transformations like FFT, DCT, *etc*, in such a way that the resulting presentation of signal is sparse (Shekhar, Patel, and Chellappa 2014). In recent years, some ADL methods have been developed. Ravishankar and Bresler (Ravishankar and Bresler 2013) proposed well-conditioned square transformations for image denoising, while Shekhar *et al.* (Shekhar, Patel, and Chellappa 2014) enhanced this method by imposing a full-rank constraint on the analysis dictionary. Rubinstein and Elad (Rubinstein and Elad 2014) imposed a hard thresholding operator on the analysis codes leading to sparse representations for synthesis reconstruction. Gu *et al.* (Gu et al. 2014) combined analysis class-specific dictionary and synthesis class-specific dictionary for classification.

These aforementioned methods benefit from the simpler optimization for training and the higher speed for testing in ADL. However, to the best of our knowledge, there are few works to address the discriminability of ADL. This may be because of its poor discriminability for pattern classification tasks. Inspired by the significant attempts of synthesis dictionary learning, we further integrate structure preserving and discriminative characters into the basic analysis model. We simply utilize the classical k Nearest Neighbor (kNN) classifier that performs exceptionally well without training effort. Nevertheless, it critically relies on the local topological structures (known as topology property in (Luo et al. 2011)) as well as the discriminability of data. Hence, it is extremely important to simultaneously exploit the underlying geometrical structures and discriminability of data when applying kNN classifier to DL-based classification tasks.

Main contributions of this paper are as follows:

- We explicitly introduce the discriminative information into the analysis dictionary learning framework via a code consistent term. Then, the learned analysis dictionary can exploit the discriminability of data instead of merely well representing data.
- We employ triplet constraints to capture the underlying discriminative local structures of data, resulting in a novel local topology preserving loss function. This loss function can preserve the relative neighborhood proximities in a supervised manner.
- We utilize correntropy induced metric (CIM) as a robust measure to handle outliers and noise. Consequently, we develop an alternating optimization algorithm based on the alternate search strategy and half-quadratic (HQ) minimization.
- Extensive comparison experiments well validate the encouraging gain in pattern classification from our method, and demonstrate that analysis models can outperform synthesis models in pattern classification tasks if discriminative information is well treated.

2 Preliminaries

2.1 Notation Summary

Bold uppercase letters $(\mathbf{U}, \mathbf{V}, \cdots)$ stand for matrices. Bold lowercase letters $(\mathbf{u}, \mathbf{v}, \cdots)$ are vectors, while lowercase letters $(u, v, \cdots)$ are scalars. $Tr\left(\mathbf{U}\right)$, $\mathbf{U}^{-1}$ and $\mathbf{U}^T$ denote the trace, inverse and transpose of $\mathbf{U}$, respectively. $\mathbf{U}_{i\cdot}$ is the i^{th} row of $\mathbf{U}$, while $\mathbf{U}_{\cdot j}$ presents the j^{th} column of $\mathbf{U}$. $\mathbf{U}_{ij}$ means the j^{th} element in the i^{th} row of $\mathbf{U}$. $\|\mathbf{U}\|_F$ and $\|\mathbf{U}\|_0$ denote the Frobenius norm ($\sqrt{\sum_{i,j}\mathbf{U}_{ij}^2}$) and l_0 norm (number of nonzero entries), respectively. $\mathbf{U}\odot\mathbf{V}$ is the Hadamard product (element-wise multiplication) of two matrices with identical size. Moreover, $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{I}$ denote the all zeros, all ones and identity matrix with appropriate sizes, respectively.

2.2 Dictionary Learning (DL)

$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n] \in \mathbb{R}^{m_1 \times n}$ denotes the original data matrix. The core idea of DL is to learn an optimized dictionary which can effectively represent each sample $\mathbf{y}_i \in \mathbb{R}^{m_1}$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m_2 \times n}$ be the coding coefficients of $\mathbf{Y}$ over the learned dictionary.

Synthesis dictionary learning: Based on classical synthesis sparse model, most existing DL methods aim to learn a synthesis dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_{m_2}] \in \mathbb{R}^{m_1 \times m_2}$ by solving

$$\begin{array}{ll} \min\limits_{\mathbf{D},\mathbf{X}} & \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \\ s.t. & \mathbf{D} \in \mathcal{D}, \\ & \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n \end{array} \tag{1}$$

where $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ stands for the reconstruction error. T_0 is a positive integer that controls the sparsity level. $\mathcal{D}$ is a set of constraints on $\mathbf{D}$ for a well-regularized solution.

In pattern classification applications, there exist various specific regularizations added to the objective function, *e.g.*, structured incoherence of dictionary (Ramirez, Sprechmann, and Sapiro 2010), transform-invariance of dictionary (Zhang et al. 2014), joint dictionary learning and subspace clustering (Zhang, He, and Davis 2014).

Analysis dictionary learning: As a dual analysis viewpoint of the commonly used synthesis dictionary learning, analysis dictionary learning (ADL) gives an intuitive explanation like feature transformation (*e.g.* DWT). It aims to learn an analysis dictionary $\mathbf{\Omega} \in \mathbb{R}^{m_2 \times m_1}$ by solving

$$\begin{array}{ll} \min\limits_{\mathbf{\Omega},\mathbf{X}} & \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \\ s.t. & \mathbf{\Omega} \in \mathcal{W}, \\ & \|\mathbf{x}_i\|_0 \le T_0,\ i = 1, 2, \cdots, n \end{array} \tag{2}$$

where $\mathcal{W}$ is a set of constraints on $\mathbf{\Omega}$ to make the solution non-trivial. As indicated in (Shekhar, Patel, and Chellappa 2014), $\mathcal{W}$ can be matrices with either relatively small Frobenius norm or unity row-wise norm.

However, this model has poor discriminability for pattern classification. Inspired by the meaningful attempts of conventional synthesis DL methods for classification tasks, we integrate two significant functions with (2) to simultaneously exploit the discriminative geometrical structures of $\mathbf{Y}$ and promote the discriminability of $\mathbf{X}$. We postpone the detailed discussion of our design until next section.

2.3 Correntropy

Different from mean square error (MSE), a global similarity measure, correntropy is a local measure between two variables, which is more robust to outliers. Correntropy is directly related to Renyi's quadratic entropy in which Parzen windowing (a non-parametric estimation method) is employed to estimate the data's probability distribution.

Non-Parametric Renyi's Entropy: Renyi's quadratic entropy is often used to measure how regular a data set is. Suppose that all $\mathbf{y}_i$s in the aforementioned data set $\mathbf{Y}$ are independently and identically drawn from the probability density function $p(\mathbf{y})$. A non-parametric estimator of $\mathbf{Y}$'s Renyi's quadratic entropy can be calculated as

$$H_R\left(\mathbf{Y}\right) = -\log \sum_i \sum_j K_\sigma\left(\mathbf{y}_i, \mathbf{y}_j\right), \tag{3}$$

where $K_\sigma\left(\mathbf{y}_i, \mathbf{y}_j\right) = \exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2/\sigma^2\right)$ is Gaussian kernel density estimation in the Parzen windowing method,

Following similar arguments, we compute Renyi's cross-entropy between two sets $\mathbf{Y}$ and $\mathbf{Y}'$ (Yuan and Hu 2009).

$$H_R\left(\mathbf{Y}; \mathbf{Y}'\right) = -\log \sum_i \sum_j K_\sigma\left(\mathbf{y}_i, \mathbf{y}'_j\right) \tag{4}$$

Correntropy Induced Metric (CIM): Based on the above concepts and a finite number of data $\{(\mathbf{y}_i, \mathbf{y}'_i)\}_{i=1}^n$, the correntropy of $(\mathbf{Y}, \mathbf{Y}')$ is defined as

$$\hat{V}_\sigma\left(\mathbf{Y}, \mathbf{Y}'\right) = \frac{1}{n}\sum_{i=1}^n K_\sigma\left(\mathbf{y}_i, \mathbf{y}'_i\right) \tag{5}$$

It is obvious that the value of correntropy is primarily decided by the Gaussian kernel function along the line $\mathbf{Y} = \mathbf{Y}'$.

Liu *et al.* (Liu, Pokharel, and Principe 2007) extended the concept of correntropy for Correntropy Induced Metric (CIM), which is a general similarity measure between any two vectors $(\mathbf{u}, \mathbf{v})$ of the same length. It is defined as

$$\begin{aligned} CIM\left(\mathbf{u}, \mathbf{v}\right) &= \left[K_\sigma\left(\mathbf{0}\right) - K_\sigma\left(\mathbf{u}, \mathbf{v}\right)\right]^{1/2} \\ &= \left[1 - \exp\left(-\left\|\mathbf{u} - \mathbf{v}\right\|_2^2/\sigma^2\right)\right]^{1/2} \end{aligned} \tag{6}$$

CIM is proved to obey all the properties for a distance metric, such as nonnegativity, identity of indiscernibles, symmetry and triangle inequality. Hence, correntropy induced metric can replace other commonly used distance metrics.

The robustness and effectiveness of correntropy have been verified in principal component analysis (He et al. 2011), feature selection (He et al. 2012), subspace clustering (Lu et al. 2013), and sparse representation (He et al. 2014). We employ this concept in our paper since it contributes to classification by well controlling outliers. To the best of our knowledge, there is no such work for analysis dictionary based pattern classification tasks.

3 The Proposed DADL Method

In this section, we introduce our proposed method, which incorporates the original data's discriminative information and local topological property into a unified analysis dictionary learning framework.

3.1 Code Consistent Term

To introduce the discriminative information into analysis dictionary learning, we design a sufficiently sparse matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n] \in \mathbb{R}^{m_2 \times n}$ as target codes. Each column of $\mathbf{H}$ is a sparse vector $\mathbf{h}_i$, which acts as the desired form of sparse code $\mathbf{x}_i$ and also indicates the class label of $\mathbf{y}_i$. Then, a code consistent term $\|\mathbf{X} - \mathbf{H}\|_F^2$ is added to (2), which can fully exploit the discrimination embedded in data. This code consistent term enforces samples from the same category to have similar sparse codes.

There are two main ways used to generate target codes, *i.e.* unsupervised and supervised manners. For the unsupervised design approach, the sparse codes for each class should be unique. Random binary codes, Hadamard codes (Yang et al. 2015), unsupervised version of Iterative Quantization are good alternatives. For the supervised design approach, the target output of multi-class linear regression is a nice choice, which is also the spectral matrix in linear discriminant analysis (LDA). In the spectral matrix, code vectors for any class have a similar form: $[0, ..., 0, 1, , 0..., 0]^T$, whose non-zero position indicates the category which is obviously sparse. However, this kind of target codes have a unified length same as the total number of classes. For analysis dictionary learning, the length of each sparse code is often larger than the number of classes. To solve this problem, we skillfully employ the Kronecker product to generate high-dimensional codes. The Kronecker product of the original spectral codes and an all ones vector with appropriate size is named Kron-form spectral codes.

In this paper, we focus on the discriminability of analysis dictionary learning rather than how to choose better target codes. Therefore, for simplicity, we generate longer sparse target codes by concatenating the Kron-form spectral codes to the sequency Walsh-ordered Hadamard codes.

3.2 Local Topology Preserving Loss Function

Local topology property describes data's local structures. Besides the neighborhood relationships, it emphasizes the ranking information of each data point's neighbors. The relative neighborhood proximities as well as the discriminability of data play a vital role in kNN classification. Different from conventional unsupervised similarity preserving loss functions based on pairwise/doublet constraints, we propose a local topology preserving loss function via triplet constraints in a supervised manner.

Definition 1. (Luo et al. 2011) Let $(\mathbf{y}_i, \mathbf{y}_u, \mathbf{y}_v)$ be a triplet comprised of $\mathbf{y}_i$ and its neighbors $\mathbf{y}_u$ and $\mathbf{y}_v$. Their corresponding codes also form a triplet $(\mathbf{x}_i, \mathbf{x}_u, \mathbf{x}_v)$. $dist(\cdot, \cdot)$ is a function that returns the pairwise distance of inputs. Then, a coding process is called *local topology preserving* when the following condition holds: if $dist(\mathbf{y}_i, \mathbf{y}_u) \leq dist(\mathbf{y}_i, \mathbf{y}_v)$, then $dist(\mathbf{x}_i, \mathbf{x}_u) \leq dist(\mathbf{x}_i, \mathbf{x}_v)$. □

Based on Definition 1, determining appropriate $\{\mathbf{x}_u, \mathbf{x}_v\}$ for $\mathbf{x}_i$ is identical to optimize

$$\max_{\mathbf{x}_u, \mathbf{x}_v} \mathbf{A}_i(u,v)\left[dist(\mathbf{x}_i, \mathbf{x}_u) - dist(\mathbf{x}_i, \mathbf{x}_v)\right], \qquad (7)$$

where $\mathbf{A}_i$ is an antisymmetric matrix whose $(u,v)^{th}$ element equals $dist(\mathbf{y}_i, \mathbf{y}_u) - dist(\mathbf{y}_i, \mathbf{y}_v)$.

However, (7) is an unsupervised type. In consideration of each sample's category, we further develop a supervised type loss by replacing $\mathbf{A}_i$ with $\mathbf{A'}_i$ in (7). The $(u,v)^{th}$ element of $\mathbf{A'}_i$ is defined as

$$\mathbf{A'}_i(u,v) \triangleq \begin{cases} -\mathbf{A}_i(u,v)\, sign\left[\mathbf{A}_i(u,v)\right] & , \mathbf{h}_i = \mathbf{h}_u \neq \mathbf{h}_v \\ \mathbf{A}_i(u,v)\, sign\left[\mathbf{A}_i(u,v)\right] & , \mathbf{h}_i = \mathbf{h}_v \neq \mathbf{h}_u \\ \mathbf{A}_i(u,v) & , otherwise \end{cases} \qquad (8)$$

where $sign(a) = \begin{cases} -1 & , a < 0 \\ 0 & , a = 0 \\ +1 & , a > 0 \end{cases}$ is the sign function. It is obvious that $\mathbf{A'}_i$ is also an antisymmetric matrix. We obtain the supervised local topology preserving loss function

$$\max_{\mathbf{X}} \sum_{i=1}^{n} \sum_{u=1}^{n} \sum_{v=1}^{n} \mathbf{A'}_i(u,v)\left[dist(\mathbf{x}_i, \mathbf{x}_u) - dist(\mathbf{x}_i, \mathbf{x}_v)\right]. \qquad (9)$$

Proposition 1. *Let* $\mathbf{W} \in \mathbb{R}^{n \times n}$ *be a weighting matrix whose* $(i,j)^{th}$ *element equals* $\sum_{u=1}^{n} \mathbf{A'}_i(u,j)$. *Objective (9) is equivalent to* $\min_{\mathbf{X}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{W}_{ij} dist(\mathbf{x}_i, \mathbf{x}_j)$.

Proof. Recall that $\mathbf{A'}_i$ is an antisymmetric matrix, so we obtain $\mathbf{A'}_i(u,v) = -\mathbf{A'}_i(v,u)$. Then (9) is equivalent to

$$\max_{\mathbf{X}} \left\{ \begin{array}{l} -\sum\limits_{i=1}^{n} \sum\limits_{u=1}^{n} \sum\limits_{v=1}^{n} \mathbf{A'}_i(v,u)\, dist(\mathbf{x}_i, \mathbf{x}_u) \\ -\sum\limits_{i=1}^{n} \sum\limits_{v=1}^{n} \sum\limits_{u=1}^{n} \mathbf{A'}_i(u,v)\, dist(\mathbf{x}_i, \mathbf{x}_v) \end{array} \right\}, \qquad (10)$$

which can also be written as

$$\max_{\mathbf{X}} \left\{ \begin{array}{l} -\sum\limits_{i=1}^{n} \sum\limits_{u=1}^{n} \mathbf{W}_{iu} dist(\mathbf{x}_i, \mathbf{x}_u) \\ -\sum\limits_{i=1}^{n} \sum\limits_{v=1}^{n} \mathbf{W}_{iv} dist(\mathbf{x}_i, \mathbf{x}_v) \end{array} \right\}. \qquad (11)$$

The ultimate form $\min_{\mathbf{X}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{W}_{ij} dist(\mathbf{x}_i, \mathbf{x}_j)$ can be easily derived from (11). □

To simultaneously preserve neighborhood ranking information as well as neighborhood relationship in a supervised manner, we employ (12) to calculate $\mathbf{W}_{ij}$.

$$\mathbf{W}_{ij} = \begin{cases} \sum\limits_{\mathbf{y}_u \in \mathcal{N}_i} \mathbf{A'}_i(u,j) & , \mathbf{y}_j \in \mathcal{N}_i \\ 0 & , otherwise \end{cases} \qquad (12)$$

where $\mathcal{N}_i$ is a set containing the k nearest neighbors of $\mathbf{y}_i$. In practice, each non-zero $\mathbf{W}_{ij}$ is normalized to $[0,1]$ by row.

Considering the important role of $\mathbf{\Omega}$ in coding process, we substitute $\mathbf{x}_i$ with $\mathbf{\Omega}\mathbf{y}_i$ in the loss function (9) so that $\mathbf{\Omega}$ can be directly learned, which is demonstrated to be efficient and effective in our experiments.

3.3 Correntropy Induced Objective Function

As indicated in (Chen and Principe 2012; Chen et al. 2014; 2015), correntropy can contribute to pattern classification by well controlling outliers. Hence, we apply CIM (6) as a robust metric and obtain the following correntropy induced objective function

$$\begin{array}{ll}\min\limits_{\mathbf{\Omega},\mathbf{X}} & J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \\ s.t. & \mathbf{\Omega} \in \mathcal{W}, \\ & \left\|\mathbf{x}_i\right\|_0 \le T_0,\ \forall i \end{array} \tag{13}$$

where

$$\left\{\begin{array}{l} J_0 = \sum\limits_{i=1}^{n} \left\{1 - \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2^2}{\sigma^2}\right)\right\} \\ J_1 = \sum\limits_{i=1}^{n} \left\{1 - \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{h}_i\|_2^2}{\sigma^2}\right)\right\} \\ J_2 = \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \left\{\mathbf{W}_{ij}\left[1 - \exp\left(-\frac{\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\|_2^2}{\sigma^2}\right)\right]\right\} \end{array}\right. , \tag{14}$$

λ_1 and λ_2 are scalar constants which control the relative importance of corresponding terms.

4 Optimization

4.1 Half-quadratic Technique

The problem (13) is not convex, so it is difficult to optimize directly. Fortunately, the half-quadratic (HQ) technique can be employed to optimize this non-convex function by alternately minimizing its augmented function. According to the conjugate function theory and HQ theory (Nikolova and Ng 2005), we have

Lemma 1. *Suppose that $f(z)$ is a function which satisfies the conditions listed in (Nikolova and Ng 2005), then for a fixed z, there exists a dual potential function $\varphi(\cdot)$, such that*

$$f(z) = \inf_{p \in \mathbb{R}} \left\{pz^2 + \varphi(p)\right\} \tag{15}$$

where p is an auxiliary variable determined by the minimizer function $\delta(z)$ w.r.t. $f(z)$.

As indicated in (Zhang et al. 2013; He, Tan, and Wang 2014), $\delta(z) = \exp\left(-\frac{z^2}{\sigma^2}\right)$ when $f(z) = 1 - \exp\left(-\frac{z^2}{\sigma^2}\right)$. That is to say, the infimum of $f(z)$ for a fixed z can be reached at $p = \delta(z)$.

According to Lemma 1, the augmented function $\hat{J}$ of (13) takes the following form

$$\begin{array}{ll}\min\limits_{\mathbf{\Omega},\mathbf{X},\mathbf{P},\mathbf{Q},\mathbf{R}} & \hat{J} = \hat{J}_0 + \lambda_1 \hat{J}_1 + \lambda_2 \hat{J}_2 \\ s.t. & \mathbf{\Omega} \in \mathcal{W}, \\ & \left\|\mathbf{x}_i\right\|_0 \le T_0,\ \forall i \end{array} \tag{16}$$

where

$$\left\{\begin{array}{l} \hat{J}_0 = \sum\limits_{i=1}^{n} \left\{\mathbf{P}_{ii}\frac{\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2^2}{\sigma^2} + \phi_i(\mathbf{P}_{ii})\right\} \\ \hat{J}_1 = \sum\limits_{i=1}^{n} \left\{\mathbf{Q}_{ii}\frac{\|\mathbf{x}_i - \mathbf{h}_i\|_2^2}{\sigma^2} + \varphi_i(\mathbf{Q}_{ii})\right\} \\ \hat{J}_2 = \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \left\{\mathbf{W}_{ij}\mathbf{R}_{ij}\frac{\left\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\right\|_2^2}{\sigma^2} + \mathbf{W}_{ij}\psi_{ij}(\mathbf{R}_{ij})\right\} \end{array}\right. \tag{17}$$

The $n \times n$ matrices $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ store the auxiliary variables introduced by HQ. Note that $\mathbf{P}$ and $\mathbf{Q}$ are diagonal. $\{\phi_i\}_{i=1}^n$, $\{\varphi_{ij}\}_{i=1}^n$ and $\{\psi_{ij}\}_{i,j=1}^n$ are conjugate functions.

4.2 Optimization Procedure

Based on the HQ optimization theory, $\hat{J}(\mathbf{\Omega}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \mathbf{R})$ can be alternately minimized as follows:

1) Update the analysis dictionary and sparse codes.

$$\left(\mathbf{\Omega}^{t+1}, \mathbf{X}^{t+1}\right) = \begin{array}{ll}\arg\min\limits_{\mathbf{\Omega},\mathbf{X}} & F = F_0 + \lambda_1 F_1 + \lambda_2 F_2 \\ s.t. & \mathbf{\Omega} \in \mathcal{W}, \\ & \left\|\mathbf{x}_i\right\|_0 \le T_0,\ \forall i \end{array} \tag{18}$$

where

$$\left\{\begin{array}{l} F_0 = Tr\left(\left(\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\right)\mathbf{P}^t\left(\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\right)^T\right) \\ F_1 = Tr\left(\left(\mathbf{X} - \mathbf{H}\right)\mathbf{Q}^t\left(\mathbf{X} - \mathbf{H}\right)^T\right) \\ F_2 = Tr\left(\mathbf{\Omega}\mathbf{Y}\mathbf{L}^{t+1}\mathbf{Y}^T\mathbf{\Omega}^T\right) \end{array}\right. . \tag{19}$$

$\mathbf{L}^{t+1}$ is the Laplacian matrix[1] of the weighting matrix $\mathbf{W}^{t+1}$ in the $(t+1)^{th}$ iteration. The weighting matrix is updated as $\mathbf{W}^{t+1} = \mathbf{W}^t \odot \mathbf{R}^t$. Let $\mathbf{C}^{t+1}$ be a diagonal matrix whose $(i,i)^{th}$ element equals $\sum_{j=1}^n \frac{\mathbf{W}_{ij}^{t+1} + \mathbf{W}_{ji}^{t+1}}{2}$. We define the Laplacian matrix $\mathbf{L}^{t+1} \triangleq \mathbf{C}^{t+1} - \frac{\mathbf{W}^{t+1} + \left(\mathbf{W}^{t+1}\right)^T}{2}$.

To fast solve (18), we prefer the set of constraints $\mathcal{W}$ to be matrices with relatively small Frobenius norm. Then, the alternate search strategy can be employed to alternatively minimize (18) with respect to one variable while fixing the other one. To update $\mathbf{\Omega} \in \mathbb{R}^{m_2 \times m_1}$ with fixed $\mathbf{X}$, we solve

$$\min_{\mathbf{\Omega}} Tr\left\{\begin{array}{l} -2\mathbf{X}\mathbf{P}^t\mathbf{Y}^T\mathbf{\Omega}^T + \lambda_3\mathbf{\Omega}\mathbf{\Omega}^T \\ +\mathbf{\Omega}\mathbf{Y}\left(\mathbf{P}^t + \lambda_2\mathbf{L}^t\right)\mathbf{Y}^T\mathbf{\Omega}^T \end{array}\right\}, \tag{20}$$

where λ_3 is the Lagrange multiplier for $\|\mathbf{\Omega}\|_F^2$. The analytical solution is computed by setting its first derivative to zero: $\mathbf{\Omega} = \mathbf{X}\mathbf{P}^t\mathbf{Y}^T\left[\mathbf{Y}\left(\mathbf{P}^t + \lambda_2\mathbf{L}^t\right)\mathbf{Y}^T + \lambda_3\mathbf{I}\right]^{-1}$. To update $\mathbf{X} \in \mathbb{R}^{m_2 \times n}$ with fixed $\mathbf{\Omega}$, we solve (21) for each column.

$$\begin{array}{ll}\min\limits_{\mathbf{x}_i} & \left\|\mathbf{x}_i - \frac{\mathbf{P}_{ii}^t\mathbf{\Omega}\mathbf{y}_i + \lambda_1\mathbf{Q}_{ii}^t\mathbf{h}_i}{\mathbf{P}_{ii}^t + \lambda_1\mathbf{Q}_{ii}^t}\right\|_2^2 \\ s.t. & \left\|\mathbf{x}_i\right\|_0 \le T_0 \end{array} \tag{21}$$

The analytical solution is obtained by applying hard thresholding operation: setting the smallest $m_2 - T_0$ elements (in magnitude) of $\frac{\mathbf{P}_{ii}^t\mathbf{\Omega}\mathbf{y}_i + \lambda_1\mathbf{Q}_{ii}^t\mathbf{h}_i}{\mathbf{P}_{ii}^t + \lambda_1\mathbf{Q}_{ii}^t}$ to 0. The update for $\mathbf{X}$ can be efficiently implemented in parallel.

2) Update auxiliary variables.

$$\mathbf{P}_{ii}^{t+1} = \exp\left(-\frac{\left\|\mathbf{x}_i^{t+1} - \mathbf{\Omega}^{t+1}\mathbf{y}_i\right\|_2^2}{\sigma^2}\right) \tag{22}$$

$$\mathbf{Q}_{ii}^{t+1} = \exp\left(-\frac{\left\|\mathbf{x}_i^{t+1} - \mathbf{h}_i\right\|_2^2}{\sigma^2}\right) \tag{23}$$

$$\mathbf{R}_{ij}^{t+1} = \exp\left(-\frac{\left\|\mathbf{\Omega}^{t+1}\mathbf{y}_i - \mathbf{\Omega}^{t+1}\mathbf{y}_j\right\|_2^2}{\sigma^2}\right) \tag{24}$$

As summarized in Algorithm 1, the above update steps are alternatively minimized until convergence.

[1] The normalized Laplacian matrix is often used in practice.

Algorithm 1 Discriminative Analysis Dictionary Learning

```
Input:
    Training data $\mathbf{Y}$ and corresponding target codes $\mathbf{H}$;
    Number of each sample's nearest neighbors $k$;
    Regularization parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$;
    Gaussian kernel parameter $\sigma$.
Output:
    The analysis dictionary $\mathbf{\Omega}$.
Set $\mathbf{X}^{(0)} = \mathbf{H}$, $\mathbf{P}^{(0)} = \mathbf{I}$, $\mathbf{Q}^{(0)} = \mathbf{I}$, $\mathbf{R}^{(0)} = \mathbf{1}$ for initialization, $t = 0$;
while not convergence do
    $t \leftarrow t + 1$;
    Compute the weighting matrix $\mathbf{W}^{(t)}$ and its Laplacian matrix $\mathbf{L}^{(t)}$;
    Update $\mathbf{\Omega}^{(t)}$ by solving (20);
    Update $\mathbf{X}^{(t)}$ by solving (21) in parallel;
    Update $\mathbf{P}^{(t)}$, $\mathbf{Q}^{(t)}$ and $\mathbf{R}^{(t)}$ via (22), (23) and (24);
end while
```

4.3 Convergence Analysis

According to Lemma 1, when $\mathbf{\Omega}$ and $\mathbf{X}$ are fixed, the following equation holds:

$$J(\mathbf{\Omega}, \mathbf{X}) = \inf_{\mathbf{P},\mathbf{Q},\mathbf{R}} \hat{J}(\mathbf{\Omega}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \mathbf{R}). \tag{25}$$

It follows that

$$\min_{\mathbf{\Omega},\mathbf{X}} J(\mathbf{\Omega}, \mathbf{X}) = \min_{\mathbf{\Omega},\mathbf{X},\mathbf{P},\mathbf{Q},\mathbf{R}} \hat{J}(\mathbf{\Omega}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \mathbf{R}). \tag{26}$$

Therefore, minimizing $J(\mathbf{\Omega}, \mathbf{X})$ is equivalent to minimizing the augmented function $\hat{J}(\mathbf{\Omega}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \mathbf{R})$ on the enlarged domain. According to the properties of half-quadratic minimization (Nikolova and Ng 2005; Yuan and Hu 2009; He et al. 2014) and alternate search strategy, we have

$$\begin{array}{ll} & \hat{J}\left(\mathbf{\Omega}^{t+1}, \mathbf{X}^{t+1}, \mathbf{P}^{t+1}, \mathbf{Q}^{t+1}, \mathbf{R}^{t+1}\right) \\ \leq & \hat{J}\left(\mathbf{\Omega}^{t+1}, \mathbf{X}^{t+1}, \mathbf{P}^{t}, \mathbf{Q}^{t}, \mathbf{R}^{t}\right) \\ \leq & \hat{J}\left(\mathbf{\Omega}^{t}, \mathbf{X}^{t}, \mathbf{P}^{t}, \mathbf{Q}^{t}, \mathbf{R}^{t}\right) \end{array}. \tag{27}$$

The objective function is non-increasing at each alternative minimization step.

What's more, according to the property of correntropy (Liu, Pokharel, and Principe 2007), the objective function $J(\mathbf{\Omega}, \mathbf{X})$ in (13) is bounded below, and thus by (26) we obtain that $\hat{J}(\mathbf{\Omega}, \mathbf{X}, \mathbf{P}, \mathbf{Q}, \mathbf{R})$ is also bounded. Consequently, we can conclude that $\hat{J}(\mathbf{\Omega}^t, \mathbf{X}^t, \mathbf{P}^t, \mathbf{Q}^t, \mathbf{R}^t)$ decreases step by step until Algorithm 1 converges.

5 Experiments

5.1 Datasets

We demonstrate the performance of our proposed method on five benchmark databases: two face datasets (YaleB and AR), and one object categorization dataset (Caltech 101), and one scene categorization dataset (Scene 15), and one action recognition dataset (UCF 50). We use the features of these databases provided by Jiang[2] and Corso[3].

The Extended Yale face database B (hereafter referred to as YaleB) (Georghiades, Belhumeur, and Kriegman 2001) includes 2,414 face images of 38 persons under 64 illumination conditions, which is challenging due to plentiful expressions and varying illumination conditions. All the original images are cropped to 192×168 pixels and then projected onto 504-dimensional vectors with a randomly generated matrix to obtain random-face features. We randomly select 32 images each person for training and the rest for testing.

The AR face database (Martinez and Benavente 1998) contains a number of color face images from 126 people. Each person has 26 frontal face images which are taken during two sessions. This database includes frontal views of faces with different facial expressions, lighting conditions, and occlusion conditions (sunglasses and scarves). All the images are cropped and scaled to 165×120. Following the standard evaluation protocol, a subset consisting of 2,600 images from 50 males and 50 females is obtained. The features used here are 540-dimensional random-face features. We randomly select 20 images each human subject for training and the other 6 images for testing.

The Caltech 101 database (Li, Fergus, and Perona 2007) is comprised of 9,144 images from 102 classes. Each category has 31 to 800 images. We use the standard Bag-of-Features (BoF) + Spatial Pyramid Matching (SPM) frame (Lazebnik, Schmid, and Ponce 2006) for feature extraction. Dense SIFT descriptors are first extracted from patches of size 16×16 which are sampled by a grid with a 6-pixel step size. Then, we compute the SPM features with 1×1, 2×2, and 4×4 subregions. Vector quantization based coding method is used to extract mid-level features and high-dimensional features are obtained by max pooling. Finally, the high-dimensional features are reduced to 3,000 dimensions by Principal Component Analysis (PCA). 30 images per category are randomly selected for training and the remaining for testing.

The fifteen scene dataset (Scene 15) was first introduced in (Lazebnik, Schmid, and Ponce 2006). The number of samples in each category ranges from 200 to 400, and the average image size is around 250×300 pixels. This database contains 15 scenes, such as kitchen, bedroom, and country scenes. Similar to Caltech 101's features, the image features used here are generated by extracting dense SIFT descriptors in local regions, encoding local patch features, max pooling in spatial pyramid and being reduced to 3,000 dimensions by PCA. Following the the common experimental settings, 100 images per category are randomly chosen as training data with the rest as testing data.

The UCF 50 (Reddy and Shah 2013) is an action recognition database with 50 action categories, consisting of 6,680 realistic human action videos taken from YouTube. For all the 50 categories, the action videos are divided into 25 groups, where each group contains over 4 action clips. The action clips in the same group share some common features, such as similar viewpoint, similar background, the same person, and so on. We utilize the action bank features (Sadanand and Corso 2012) and five-fold data splitting to evaluate our method, where four folds are used for training and the remaining one fold for testing. We employ PCA to reduce the

[2] http://www.umiacs.umd.edu/∼zhuolin/projectlcksvd.html.

[3] http://www.cse.buffalo.edu/∼jcorso/r/actionbank.

Table 1: Major parameters, determined by cross-validation.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
k	7	7	5	5	7
λ_2	0.001	0.001	0.001	0.010	0.001
λ_3	0.100	0.100	4.000	1.000	0.100

Table 2: Classification accuracies (%) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
ADL+SVM	95.4	96.1	64.5	90.1	72.3
SRC	96.5	97.5	70.7	91.8	75.0
CRC	97.0	98.0	68.2	92.0	75.6
DLSI	97.0	97.5	73.1	91.7	75.4
FDDL	96.7	97.5	73.2	92.3	76.5
LC-KSVD	96.7	97.8	73.6	92.9	70.1
DPL	97.5	98.3	73.9	97.7	77.4
DADL	**97.7**	**98.7**	**74.6**	**98.3**	**78.0**

features to 5,000 dimensions.

5.2 Experiment Setup

We compare our proposed DADL method with the following methods: the baseline Analysis Dictionary Learning + Support Vector Machine (ADL+SVM) (Shekhar, Patel, and Chellappa 2014), the classical Sparse-Representation-based Classifier (SRC) (Wright et al. 2009) and Collaborative-Representation-based Classifier (CRC) (Zhang, Yang, and Feng 2011), and three state-of-the-art dictionary learning methods: Dictionary Learning with Structured Incoherence (DLSI) (Ramirez, Sprechmann, and Sapiro 2010), Fisher Discrimination Dictionary Learning (FDDL) (Yang et al. 2011), Label Consistent K-SVD (LC-KSVD) (Jiang, Lin, and Davis 2013), and the recently proposed projective Dictionary Pair Learning (DPL) (Gu et al. 2014).

For fair comparison, we follow the experimental settings in (Gu et al. 2014) for all the competing methods. We set the Gaussian kernel parameter $\sigma = 10$ and the balance weight $\lambda_1 = 10$ in all our experiments. The experimental results are insensitive to $\sigma \in [7, 13]$ and $\lambda_1 \in [10, 15]$. Since sparsity level depends on the sparse target codes $\mathbf{H}$ that is determined by information theoretic rules (refer to Section 3.1), it can be well treated. The other major parameters $(k, \lambda_2, \lambda_3)$ on each database have been tuned by cross validation. The best $(k, \lambda_2, \lambda_3)$ for each database are listed in Table 1.

5.3 Results and Analysis

Given training data and corresponding target codes, an optimized dictionary $\mathbf{\Omega}$ can be learned by Algorithm 1. Then, we can code both training and testing samples via the learned $\mathbf{\Omega}$. Finally, we treat these coding vectors as new features and employ kNN classifier to perform classification. All the experiments are repeated 20 times with different random spits of training and testing images on each dataset. Reliable results of different methods are reported in Table 2.

Table 3: Training time (s) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
DPL	5.92	15.21	180.54	56.84	652.03
DADL	4.23	11.16	121.47	36.52	330.23

Table 4: Testing time (ms) on five datasets.

	YaleB	AR	Caltech 101	Scene 15	UCF 50
DPL	0.19	0.42	1.45	1.36	1.62
DADL	0.16	0.39	1.39	1.31	1.48

By contrast, DADL achieves obviously higher accuracy than the basic ADL+SVM framework, which indicates that our proposed method significantly improves the discriminative ability of ADL. Compared with synthesis dictionary based classification methods (SRC, CRC, DLSI, FDDL, and LC-KSVD), our proposed DADL method achieves the best performance. Besides, our approach also gives a better result in comparison with the recently proposed projective Dictionary Pair Learning (DPL) method, which combines discriminative synthesis dictionary learning and ADL.

For some datasets, all the competitors achieve over 95% accuracy, so our method's improvement is not visibly big. We can observe that DPL and our DADL obviously outperform other DL methods, which from another side proves the strong vitality of analysis dictionary in classification tasks. DPL in (Gu et al. 2014) also markedly outperforms state-of-the-art DL methods in terms of faster running time. Therefore, we conduct extra experiments to further evaluate the efficiency of our method. Our experiments are run via MATLAB R2013a on a desktop PC with an Intel Core i7-3770 processor at 3.40 GHz and 16.00 GB RAM. As reported in Table 3 and 4, the less time consumption shows the superiority of our method. Thus we can safely conclude that our proposed DADL method performs better in classification than state-of-the-art dictionary learning methods.

6 Conclusion

In this paper, we propose a Discriminative Analysis Dictionary Learning (DADL) method. To make analysis dictionary learning (ADL) applicable for pattern classification tasks, a code consistent term has been introduced. Meanwhile, based on triplet constraints, a discriminative local topology preserving loss function has been developed, which simultaneously preserves neighborhood relationship as well as proximities in a supervised manner. Besides, correntropy induced metric is utilized as a robust measure to improve robustness. Based on half-quadratic (HQ) technique and alternate search strategy, we have developed an iterative method to speed up the ADL process. Experimental results on five benchmark datasets show the effectiveness of our method against state-of-the-art dictionary learning methods.

Acknowledgments

This work is funded by the National Natural Science Foundation of China (Grant No. 61402079, 61473289, 61175003, 61172109), the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the Youth Innovation Promotion Association CAS (Grant No. 2015190), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02000000), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR). We thank the anonymous reviewers and Dr. Ming Li from DLUT for their helpful comments.

References

Chen, B., and Principe, J. C. 2012. Maximum correntropy estimation is a smoothed map estimation. *IEEE SPL* 19(8):491–494.

Chen, B.; Xing, L.; Liang, J.; Zheng, N.; and Principe, J. C. 2014. Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. *IEEE SPL* 21(7):880–884.

Chen, B.; Wang, J.; Zhao, H.; Zheng, N.; and Principe, J. C. 2015. Convergence of a fixed-point algorithm under maximum correntropy criterion. *IEEE SPL* 22(10):1723–1727.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI* 23(6):643–660.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *Proc. NIPS*, 793–801.

He, R.; Hu, B.; Zheng, W.; and Kong, X. 2011. Robust principal component analysis based on maximum correntropy criterion. *IEEE TIP* 20(6):1485–1494.

He, R.; Tan, T.; Wang, L.; and Zheng, W. 2012. $l_{2,1}$ regularized correntropy for robust feature selection. In *Proc. CVPR*, 2504–2511.

He, R.; Zheng, W.; Tan, T.; and Sun, Z. 2014. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE TPAMI* 36(2):261–275.

He, R.; Tan, T.; and Wang, L. 2014. Recovery of corrupted low-rank matrix by implicit regularizers. *IEEE TPAMI* 36(4):770–783.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2013. Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE TPAMI* 35(11):2651–2664.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, volume 2, 2169–2178.

Li, F.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *CVIU* 106(1):59–70.

Liu, W.; Pokharel, P. P.; and Principe, J. C. 2007. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE TSP* 55(11):5286–5298.

Lu, C.; Tang, J.; Lin, M.; Lin, L.; Yan, S.; and Lin, Z. 2013. Correntropy induced l_2 graph for robust subspace clustering. In *Proc. ICCV*, 1801–1808.

Luo, D.; Ding, C.; Nie, F.; and Huang, H. 2011. Cauchy graph embedding. In *Proc. ICML*, 553–560.

Martinez, A., and Benavente, R. 1998. The AR face database. *CVC Tech. Rep.* 24.

Nikolova, M., and Ng, M. K. 2005. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Scientific Computing* 27(3):937–966.

Ramirez, I.; Sprechmann, P.; and Sapiro, G. 2010. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proc. CVPR*, 3501–3508.

Ravishankar, S., and Bresler, Y. 2013. Learning sparsifying transforms. *IEEE TSP* 61(5):1072–1086.

Reddy, K. K., and Shah, M. 2013. Recognizing 50 human action categories of web videos. *Mach. Vision Applicat.* 24(5):971–981.

Rubinstein, R., and Elad, M. 2014. Dictionary learning for analysis-synthesis thresholding. *IEEE TSP* 62(22):5962–5972.

Rubinstein, R.; Bruckstein, A. M.; and Elad, M. 2010. Dictionaries for sparse representation modeling. *Proc. IEEE* 98(6):1045–1057.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *Proc. CVPR*, 1234–1241.

Shekhar, S.; Patel, V. M.; and Chellappa, R. 2014. Analysis sparse coding models for image-based classification. In *Proc. ICIP*, 5207–5211.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE TPAMI* 31(2):210–227.

Yang, M.; Zhang, L.; Feng, X.; and Zhang, D. 2011. Fisher discrimination dictionary learning for sparse representation. In *Proc. ICCV*, 543–550.

Yang, S.; Luo, P.; Loy, C. C.; Shum, K. W.; and Tang, X. 2015. Deep representation learning with target coding. In *Proc. AAAI*.

Yuan, X., and Hu, B. 2009. Robust feature extraction via information theoretic learning. In *Proc. ICML*, 1193–1200.

Zhang, Y.; Sun, Z.; He, R.; and Tan, T. 2013. Robust subspace clustering via half-quadratic minimization. In *Proc. ICCV*, 3096–3103.

Zhang, S.; Zhang, M.; He, R.; and Sun, Z. 2014. Transform-invariant dictionary learning for face recognition. In *Proc. ICIP*, 348–352.

Zhang, G.; He, R.; and Davis, L. S. 2014. Jointly learning dictionaries and subspace structure for video-based face recognition. In *Proc. ACCV*, 348–352.

Zhang, L.; Yang, M.; and Feng, X. 2011. Sparse representation or collaborative representation: Which helps face recognition? In *Proc. ICCV*, 471–478.