DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-456

Class specific discriminant dictionary learning with kernels for face recognition

Bao-Di Liu*[1]*; College of Information and Control Engineering, China University of Petroleum; Qingdao, 266580, China
Yuting Wang; Department of Informatics, Karlsruhe Institute of Technology; Karlsruhe, 76131, Germany
Liangke Gui; School of Computer Science, Carnegie Mellon University; Pittsburgh, PA 15213, USA
Yu-Xiong Wang; School of Computer Science, Carnegie Mellon University; Pittsburgh, PA 15213, USA
Bin Shen; Department of Computer Science, Purdue University; West Lafayette, IN 47907 USA
Xue Li; Department of Electronic Engineering, Tsinghua University; Beijing 100084, China
Yan-Jiang Wang; College of Information and Control Engineering, China University of Petroleum; Qingdao, 266580, China

Abstract

The past few years have witnessed the impressive performance of sparse representation based classification (SRC) for visual recognition. However, the SRC technique may lead to high residual error and poor performance due that the training samples in each class contribute equally to the dictionary in the corresponding class. This inspired the emergence of class specific dictionary learning algorithm. In this paper, we propose a novel approach—class specific dictionary learning combined with linear discriminant analysis constraints in Reproducing Kernel Hilbert Space (KCSDL-LDA), which modifies and extends the conventional class specific dictionary learning (CSDL) algorithm in several aspects. First, we propose a novel class specific dictionary learning scheme that considers the weight of each sample for each class when generating the dictionary in that class. Second, we extend the novel class specific dictionary learning scheme to the Reproducing Kernel Hilbert Space, in which nonlinear structure can be extracted and represented to improve the classification accuracy. Finally, we further enhance the classification performance by combing class specific dictionary learning with linear discriminant analysis constraints in Reproducing Kernel Hilbert Spaces. Extensive experimental results on several face recognition benchmark datasets, such as Extended YaleB dataset, CMU PIE dataset and AR dataset, demonstrate the superior performance of our proposed KCSDL-LDA.

Introduction

The past few years have witnessed the impressive performance of dictionary learning for sparse representation in visual computation areas, such as image annotation [1], image inpainting [2], image classification [3], face recognition [4] and image denoising [5]. Different from traditional decomposition frameworks like Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) [6] and low-rank factorization, sparse representation is capable of generating sparse codes under over-complete bases to represent the data more adaptively and flexibly.

Face recognition, one of the successful applications of sparse representation, is a classical yet challenging research topic in computer vision and pattern recognition [7]. Effective face recognition usually involves two important stages: 1) feature extraction, 2) classifier construction and face prediction. For the first stage, Turk *et al.* performed principal component analysis (PCA) to extract Eigenfaces [8]. He *et al.* proposed Laplacianfaces [9] to preserve local information. Belhumeur *et al.* extracted Fisherfaces [10] to maximize the ratio of between-class scatter to within-class scatter. For the latter stage, Richard *et al.* introduced a nearest neighbor method [11] to predict the label of a test image using its nearest neighbors in the training samples. Tao *et al.* presented a nearest subspace method [12] to assign the label of a test image by comparing its reconstruction error for each category.

Under the nearest subspace framework, Wright *et al.* [4] described a sparse representation based classification (SRC) system and achieved an impressive performance for face recognition. Given a test sample, the sparse representation technique represents it as a sparse linear combination of the train samples. The predicted label is determined by the residual error from each class. Zhang *et al.* [13] illustrated a collaborative representation based classification (CRC) system. Similar to SRC, CRC represents a test sample as the linear combination of almost all the training samples. Moreover, Zhang *et al.* demonstrated that it was the collaborative representation rather than the sparse representation that makes the nearest subspace method powerful for classification. Overall, both SRC and CRC algorithms directly use the training samples as the dictionary for each class. This may lead to high residual error and poor performance due that the training samples in each class contribute equally to the dictionary in the corresponding class. Therefore, the emergence of class specific dictionary learning algorithm attracts the attention of many researchers. They focus on learning a dictionary enforced by some discriminative criteria that can reduce the residual error greatly and achieve a superior performance for classification tasks.

So far, existing discriminative dictionary learning approaches are mainly categorized into three types: shared dictionary learning, class specific dictionary learning and hybrid dictionary learning. In shared dictionary learning, the bases are learned with all the training samples together. The discriminative information is often embedded into the dictionary learning procedure. Mairal *et al.* learned a discriminative dictionary [14] with a linear classifier of coding coefficients. Liu *et al.* embedded the linear discriminant analysis [15] into the dictionary. Zhang *et al.* obtained a discriminative dictionary by integrated the label information [16] into the dictionary learning. The shared dictionary learning approaches usually lead to a small-sized dictionary and

[1] thu.liubaodi@gmail.com

the discriminative information (i.e., the label information corresponding to coding coefficients) is embedded into the dictionary learning framework. In class specific dictionary learning, each basis is only corresponds to a single class so that the class specific reconstruction error could be used for classification. Yang *et al.* first elaborated a class specific dictionary learning algorithm [17] for face recognition. Wang *et al.* detailed the mutual incoherence information [18] to promote class specific dictionary learning in action recognition. Yang *et al.* improved the class specific dictionary learning algorithm with the fisher discriminative information [19]. Liu *et al.* depicted a self-explanatory sparse representation based dictionary learning [20] to enhance the interpretation of the class specific based dictionary learning algorithm. Liu *et al.* listed class specific centralized dictionary learning algorithm [21] to make the sparse codes in the same class centralized. The class specific dictionary learning approaches usually focus on the classifier construction aspect since each basis vector is fixed to a single class label. In hybrid dictionary learning, the shared basis vectors and class specific basis vectors are learned simultaneously. Zhou *et al.* learned a hybrid dictionary [22] with fisher regularization on the coding coefficient. Gao *et al.* learned a shared dictionary [23] to encode common visual patterns and a class specific dictionary to encode subtle visual differences among different categories for fine-grained image representation. Liu *et al.* showed a hierarchical dictionary learning method [24] to produce a shared dictionary and a cluster specific dictionary. In spite of the demonstrated performance of hybrid dictionary learning, it is still a challenge to balance the shared dictionary and the class specific dictionary.

Although the methods mentioned above achieved superior performance in visual recognition, works of dictionary learning usually operate in the original Euclidean space, which cannot capture nonlinear structures hidden in data. Meanwhile, face images often have intrinsic nonlinear similarity measures. A classical way to deal with this is to adopt the "*kernel trick*" [25], which maps the features into high dimensional feature space to make features of different categories more linearly separable. With the introduction of kernel techniques, the learned dictionary becomes versatile. Wu *et al.* learned a dictionary in the histogram intersection kernel (HIK) space [26], while Gemert *et al.* learned it in the Gaussian radial basis function (RBF) kernel space [27]. Liu *et al.* explained a self-explanatory sparse representation [28] for image classification and extended the dictionary learning to the arbitrary kernel space.

Motivated by the higher performance of class specific dictionary learning and dictionary learning in the kernel space, we propose a novel approach to combine the class specific dictionary learning with linear discriminant analysis constraints in Reproducing Kernel Hilbert Space (KCSDL-LDA), which is considered as an extension and improvement of the conventional class specific dictionary learning (CSDL) algorithm. **The main contribution** is listed in four aspects:

- We propose a novel class specific dictionary learning scheme in Reproducing Kernel Hilbert Spaces that considers the weight of each sample in correspondence class when generating the dictionary in the kernel space.
- We propose class specific dictionary learning combined with linear discriminant analysis constraints in Reproducing Kernel Hilbert Spaces for sparse representation based classification.
- We use the Coordinate descent and Lagrange multipliers to efficiently solve the corresponding optimization problems.
- We show that our proposed KCSDL-LDA algorithm achieves superior performance in several benchmark datasets of face recognition tasks to other classical face recognition algorithms.

Overview of SRC and CRC

Sparse representation and collaborative representation algorithms can be considered as methods of rearranging the structure of the original data in order to make the representation compact and discriminative under non-orthogonal bases. Hence, the data vector is represented as a linear combination of active basis vectors.

Wright *et al.* proposed the sparse representation based classification (SRC) algorithm for robust face recognition [4]. Given the training samples $X = [X^1, X^2, \cdots, X^C] \in \mathbb{R}^{D\times N}$, where $X^c \in \mathbb{R}^{D\times N_c}$ represents the training samples from the c_{th} class, C represents the number of classes, N_c represents the number of training samples in the c_{th} class ($N = \sum_{c=1}^{C} N_c$), and D represents the dimension of the samples. Supposing that $y \in \mathbb{R}^{D\times 1}$ is a test sample, the sparse representation algorithm aims to solve the following optimization problem:

$$\hat{s} = \arg min_s \left\{ \|y - Xs\|_2^2 + 2\alpha \|s\|_1 \right\}. \tag{1}$$

Here, α is the regularization parameter to control the trade-off between fitting goodness and sparseness.

Zhang *et al.* proposed the collaborative representation based classification (CRC) algorithm [13] by replacing ℓ_1 regularizer term in Eqn. (1) with ℓ_2 regularizer term as follows,

$$\hat{s} = \arg min_s \left\{ \|y - Xs\|_2^2 + \beta \|s\|_2^2 \right\}. \tag{2}$$

Here, β is the regularization parameter to control the trade-off between fitting goodness and collaborative property (i.e., multiple entries in X participating in representing the test sample).

The sparse representation or collaborative representation based classifier is to find the minimum value of the residual error for each class:

$$id(y) = \arg min_c \|y - X^c \hat{s}^c\|_2^2. \tag{3}$$

Both the SRC and CRC algorithms directly use the training samples as the dictionary and encode the test sample y as

$$y \approx XWs, \tag{4}$$

where $W \in \mathbb{R}^{N\times N}$ is an identity matrix. This means that the training samples contribute equally for constructing the dictionary $B = XW$ when representing the test sample y.

Our Approach

In this section, we propose a novel approach—class specific dictionary learning combined with linear discriminant analysis constraints in Reproducing Kernel Hilbert Space (KCSDL-LDA), which modifies and extends the conventional class specific dictionary learning (CSDL) algorithm in several aspects.

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-456

Class specific dictionary learning in the kernel space

From Eqns. (1), (2) and (4), we could observe that W is predefined as an identity matrix I. However, for different classification tasks and sample distributions, a data-driven formulation of W would be preferred. That is, to make W more adaptive, it would be of great benefit to impose that the training samples of the same class have different weights when constructing bases in the corresponding dictionary (intra-class constraint) while the training samples of the remaining classes have no contribution (inter-class constraints). Hence, the weight coefficient matrix could be generalized from an identity matrix to a block-diagonal matrix [20] as shown in Figure 1. W can be obtained effectively by dictionary learning. SRC and CRC can be thus considered as special cases of our proposed class specific dictionary learning.

$$\begin{bmatrix} W^1 & 0 & 0 & \cdots & 0 \\ 0 & W^2 & 0 & \cdots & 0 \\ 0 & 0 & W^3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & W^C \end{bmatrix}$$

Figure 1. *The learned weight coefficient matrix W for constructing dictionary*

The objective function of CSDL becomes

$$\mathscr{G}(W^1,\cdots,W^C,S^1,\cdots,S^C) = \sum_{c=1}^{C}\left\{\|X^c - X^cW^cS^c\|_F^2 + 2\alpha\sum_{n=1}^{N_c}\|S^c_{\bullet n}\|_1\right\} \quad (5)$$
$$s.t.\ \|X^cW^c_{\bullet k}\|_2^2 \le 1, \forall k=1,2,\cdots,K, \forall c=1,2,\ldots,C.$$

where $\|\bullet\|_F^2$ represents the Frobenius norm. $B_{\bullet i}$ and $B_{j\bullet}$ denote the i_{th} column and j_{th} row vectors of matrix B, respectively. W is the learned weight coefficient for constructing the dictionary and S is the corresponding sparse representation. Equation (5) can be also considered as the "dual form" of the conventional CSDL algorithm.

The conventional CSDL algorithm usually operates in Euclidean space, which could not capture the nonlinear structure when learning the dictionary. By contrast, our proposed CSDL algorithm can be easily extended to the kernel space, where the nonlinear structures are extracted and represented to enhance the classification accuracy. Suppose that there exists a feature mapping function $\phi : \mathbb{R}^D \to \mathbb{R}^t$, it maps the original feature space to the high dimensional kernel space: $\phi(X) = [\phi(X^1), \phi(X^2), \cdots, \phi(X^C)] \in \mathbb{R}^{t\times N}$. The objective function of Eqn. (5) can then be generalized to reproducing kernel Hilbert spaces as

$$\mathscr{O}(W^1,\cdots,W^C,S^1,\cdots,S^C) = \sum_{c=1}^{C}\left\{\|\phi(X^c) - \phi(X^c)W^cS^c\|_H^2 + 2\alpha\sum_{n=1}^{N_c}\|S^c_{\bullet n}\|_1\right\} \quad (6)$$
$$s.t.\ \|\phi(X^c)W^c_{\bullet k}\|_H^2 \le 1, \forall k=1,2,\cdots,K, \forall c=1,2,\ldots,C.$$

Here, the Frobenius norm has been replaced by the inner-product norm of that Hilbert space, such that $\|\phi(X)\|_H^2 = \kappa(X,X)$, with kernel function $\kappa(X_{\bullet i}, X_{\bullet j}) = \phi(X_{\bullet i})^T\phi(X_{\bullet j})$.

The KCSDL-LDA model

The conventional CSDL algorithm fails to consider the constraints crucial to the produced sparse codes. In particular, it cannot guarantee the sparse codes that will be concentrated in the same class and scattered in different class based on the learned dictionary for each class. Such within-class concentration and between-class scatter are actually beneficial to the classification. We now come to our proposed class specific dictionary learning combined with linear discriminant analysis constraints in Reproducing Kernel Hilbert Spaces for sparse representation based classification.

More precisely, the within-class concentration information is

$$\mathscr{R}(S^1,\cdots,S^C) = \sum_{c=1}^{C}\sum_{n=1}^{N_c}\left\|S^c_{\bullet n} - \frac{1}{N_c}\sum_{m=1}^{N_c}S^c_{\bullet m}\right\|_2^2. \quad (7)$$

The between-class scatter information is

$$\mathscr{T}(S^1,\cdots,S^C) = \sum_{c=1}^{C}\frac{1}{(C-1)}\sum_{d=1,d\neq c}^{C}\left\|\frac{1}{N_c}\sum_{n=1}^{N_c}S^c_{\bullet n} - \frac{1}{N_d}\sum_{m=1}^{N_d}S^d_{\bullet m}\right\|_2^2. \quad (8)$$

Intuitively, we can define $g(W,S) = \mathscr{R} - \mathscr{T}$ as our discriminant term. However, such term is non-convex and unstable. To solve this problem, we propose to add an term $\left\|\frac{1}{N_c}\sum_{n=1}^{N_c}(S^c_{\bullet n})\right\|_2^2$ into $g(W,S)$. So $g(W,S)$ is defined as

$$g(W^1,\cdots,W^C,S^1,\cdots,S^C) = \beta\{\mathscr{R} - \mathscr{T} + 2\left\|\frac{1}{N_c}\sum_{n=1}^{N_c}S^c_{\bullet n}\right\|_2^2\} \quad (9)$$

The objective function of our proposed class specific discriminant dictionary learning with kernels now becomes

$$\begin{aligned} f(W^1,\cdots,W^C,S^1,\cdots,S^C) &= \mathscr{O}(W^1,\cdots,W^C,S^1,\cdots,S^C) \\ &\quad + g(W^1,\cdots,W^C,S^1,\cdots,S^C) \quad (10) \\ s.t.\|\phi(X^c)W^c{}_{\bullet k}\|_H^2 &\le 1, \forall k=1,2,\ldots,K, \forall c=1,2,\ldots,C. \end{aligned}$$

Optimization of the objective function

In this section, we focus on solving the optimization problem for the proposed KCSDL-LDA algorithm. Specifically, similar to the optimization strategy adopted in [29, 30], it is decomposed into two subproblems via alternating minimization for learning dictionary of each class. One is an ℓ_1-norm regularized least-squares minimization subproblem with fixed W and $(S^1, S^2, \cdots, S^{c-1}, S^{c+1}, \cdots, S^C)$. The other one is an ℓ_2-norm constrained least-squares minimization subproblem with fixed S and $(W^1, W^2, \cdots, W^{c-1}, W^{c+1}, \cdots, W^C)$.

ℓ_1-ℓ_s Minimization Subproblem

With W and $(S^1, S^2, \cdots, S^{c-1}, S^{c+1}, \cdots, S^C)$ fixed, the objective function of the ℓ_1-ls minimization subproblem is cast as

$$\begin{aligned} f(S^c) &= \|\phi(X^c) - \phi(X^c)W^cS^c\|_H^2 + 2\alpha\sum_{n=1}^{N_c}\|S^c_{\bullet n}\|_1 \\ &\quad + \beta\sum_{n=1}^{N_c}\left\|S^c_{\bullet n} - \frac{1}{N_c}\sum_{m=1}^{N_c}S^c_{\bullet m}\right\|_H^2 + 2\beta\left\|\frac{1}{N_c}\sum_{n=1}^{N_c}S^c_{\bullet n}\right\|_H^2 \quad (11) \\ &\quad - \beta\frac{1}{(C-1)}\sum_{d=1,d\neq c}^{C}\left\|\frac{1}{N_c}\sum_{n=1}^{N_c}S^c_{\bullet n} - \frac{1}{N_d}\sum_{m=1}^{N_d}S^d_{\bullet m}\right\|_H^2 \end{aligned}$$

Ignoring the constant term, Eqn. (11) can be simplified as

$$\begin{aligned} f(S^c) = &\, trace\{\kappa(X^c,X^c) - 2\kappa(X^c,X^c)W^cS^c\} \\ &+ trace\left\{S^{cT}\left(W^{cT}\kappa(X^c,X^c)W^c\right)S^c\right\} + 2\alpha\sum_{n=1}^{N_c}\|S^c_{\bullet n}\|_1 \\ &+ \beta\sum_{n=1}^{N_c}\left\|\frac{N_c-1}{N_c}S^c_{\bullet n} - \frac{1}{N_c}\left(\sum_{m=1,m\neq n}^{N_c}S^c_{\bullet m}\right)\right\|_H^2 \\ &+ \beta\sum_{n=1}^{N_c}\left\|\frac{1}{N_c}S^c_{\bullet n} + \frac{1}{N_c}\left(\sum_{m=1,m\neq n}^{N_c}S^c_{\bullet m}\right)\right\|_H^2 \\ &- \beta\frac{1}{(C-1)}\sum_{d=1,d\neq c}^{C}\left\|\frac{1}{N_c}S^c_{\bullet n} + \frac{1}{N_c}\left(\sum_{m=1,m\neq n}^{N_c}S^c_{\bullet m}\right) - \frac{1}{N_d}\sum_{i=1}^{N_d}S^d_{\bullet i}\right\|_H^2. \end{aligned} \tag{12}$$

According to the solving method in [20], it is easy to infer that $f(S^c_{kn})$ reaches the minimum at the unique point

$$\begin{aligned} S^c_{kn} = &\, \frac{1}{1+\beta\left(\frac{N_c-1}{N_c}\right)^2 + \beta\left(\frac{1}{N_c}\right)^2}\min\{A_{kn} - [E\widetilde{S^c}^{kn}]_{kn}, -\alpha\} \\ &+ \frac{1}{1+\beta\left(\frac{N_c-1}{N_c}\right)^2 + \beta\left(\frac{1}{N_c}\right)^2}\max\{A_{kn} - [E\widetilde{S^c}^{kn}]_{kn}, \alpha\}, \end{aligned} \tag{13}$$

where $A_{kn} = [W^{cT}\kappa(X^c,X^c)]_{kn} + \beta\left[\frac{N_c-2}{N_c{}^2}\sum_{m=1,m\neq n}^{N}S^c_{km}\right] - \beta\left[\frac{1}{N_cN_d(C-1)}\sum_{d=1,d\neq c}^{C}\sum_{i=1}^{N}S^d_{ki}\right]$, $E = W^{cT}\kappa(X^c,X^c)W^c$, and $\widetilde{S^c}^{kn} = \begin{cases} S^c_{pq}, p\neq k \| q\neq n \\ 0, \quad p=k \& q=n \end{cases}$.

ℓ_2-ℓ_s *minimization subproblem*

With S^c and $(W^1, W^2, \cdots, W^{c-1}, W^{c+1}, \cdots, W^C)$ fixed, the objective function of the ℓ_2-ℓ_s minimization subproblem becomes

$$\begin{aligned} & f(W^c) = \|\phi(X^c) - \phi(X^c)W^cS^c\|_H^2 \\ & \quad s.t.\ \|\phi(X^c)W^c_{\bullet k}\|_H^2 \leq 1, \forall k = 1,2,\cdots,K. \end{aligned} \tag{14}$$

Here, the Lagrange multipliers are used to solve the ℓ_2-norm constrained minimization subproblem. W^c can be obtained by optimizing each column alternately. Specifically, ignoring the constant term $trace\{\kappa(X^c,X^c)\}$, the Lagrangian of Eqn. (14) is

$$\begin{aligned} \mathscr{L}(W^c,\lambda_k,\mu_k) = &-2\sum_{k=1}^{K}[S^c\kappa(X^c,X^c)]_{k\bullet}W^c_{\bullet k} \\ &+ \sum_{k=1}^{K}W^c_{\bullet k}{}^T[\kappa(X^c,X^c)W^cS^cS^{cT}]_{\bullet k} \\ &+ \lambda_k(1 - [W^{cT}\kappa(X^c,X^c)W^c]_{kk}), \end{aligned} \tag{15}$$

where λ_k is a variable.

According to the Karush-Kuhn-Tucker (KKT) conditions, the optimal solution $W^c_{\bullet k}$ should satisfy the following criteria:

$$\begin{aligned} &(a): \frac{\partial\mathscr{L}(W^c,\lambda_k)}{\partial W^c_{\bullet k}} = 0; \\ &(b): (1 - [W^{cT}\kappa(X^c,X^c)W^c]_{kk}) = 0; \\ &(c): \lambda_k > 0 \end{aligned} \tag{16}$$

Hence, the solution to $W^c_{\bullet k}$ becomes

$$W^c_{\bullet k} = \frac{S^c_{k\bullet}{}^T - [\widetilde{W^c}^kF]_{\bullet k}}{\sqrt{(S^c_{k\bullet}{}^T - [\widetilde{W^c}^kF]_{\bullet k})^T\kappa(X^c,X^c)(S^c_{k\bullet}{}^T - [\widetilde{W^c}^kF]_{\bullet k})}}, \tag{17}$$

where $F = S^cS^{cT}$ and $\widetilde{W^c}^k = \begin{cases} W^c_{\bullet p}, p\neq k \\ 0, \quad p=k \end{cases}$.

Algorithm 1 KCSDL-LDA Algorithm

```
Require: Data matrix $X \in \mathbb{R}^{D\times N}$, $\alpha$, $\beta$, $K$, and $C$;
Compute kernels $\kappa(X,X)$ on data $X$;
for $c = 1:1:C$ do
    $W^c \leftarrow rand(N_c,K), S^c \leftarrow zeros(K,N_c)$
    for $k = 1; k \leq K; k{+}{+}$ do
        $W^c_{\bullet k} = W^c_{\bullet k}/\sqrt{W^c_{\bullet k}{}^T W^c_{\bullet k}}$
    end for
end for
while not converge do
    for $c = 1:1:C$ do
        Update $S^c$ according to Eqn. (13)
        Update $W^c$ according to Eqn. (17)
    end for
    Update the objective function according to Eqn. (10)
end while
return $W$, and $S$
```

Overall algorithm

Our algorithm for KCSDL-LDA is shown in Algorithm 1.

Experimental results

In this section, we evaluated our KCSDL-LDA algorithm on three benchmark dataset, such as the Extended YaleB dataset [31], the CMU PIE dataset [32], and the AR dataset [33]. The following subsections focus on experimental settings, parameter tuning, experimental results and some discussions.

Experimental settings

For all three benchmark datasets, each face image is cropped to 32×32, pulled into a column vector, and performed a ℓ_2 normalization to form the raw feature. After that, 5 samples per class as training data and 10 samples per class as testing data are randomly selected from the datasets. To eliminate the randomness, we randomly (repeatable) split the dataset into the train set and test set 10 times, respectively. The mean value and standard deviation of the face recognition rate are recorded.

For comparison, five classical face recognition algorithms are used as baselines. They are nearest neighbor classification (N-N), collaborative representation based classification (CRC) [13], sparse representation based classification (SRC) [4], class specific dictionary learning based classification [20], and SVM [34]. For SVM, one-against-all multi-class classification strategy is adopted by LIBSVM [34].

For parameter selection, three parameters are required to tune. α is used to adjust the trade-off between the reconstruction error and the sparsity for sparse representation based class specific dictionary learning. β is used to adjust the trade-off between the reconstruction error and the discriminant information. K is the size of the dictionary for each class. In this paper, K is set to be twice of the size of the training samples per class. The detailed parameter adjustment is shown in the following subsections.

For kernel functions, we use three different kernels: linear kernel ($\kappa(x,y) = x^Ty$), the Hellinger kernel ($\kappa(x,y) = \sum_{d=1}^{D}\sqrt{x_dy_d}$), and the polynomial kernel ($\kappa(x,y) = (p + x^Ty)^q$), and . Here, we set $p = 4$ and $q = 2$.

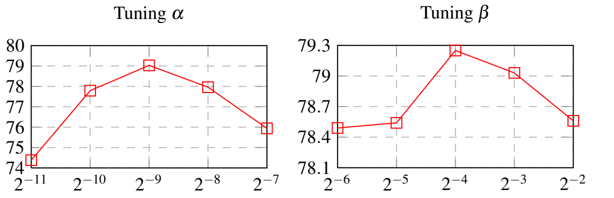

Figure 2. Parameter tuned on CMU PIE dataset with linear kernel. The left figure is for tuning α with $\beta = 2^{-4}$. The right figure is for tuning β with $\alpha = 2^{-9}$

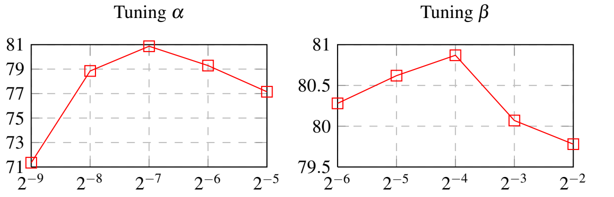

Figure 3. Parameter tuned on CMU PIE dataset with Hellinger kernel. The left figure is for tuning α with $\beta = 2^{-4}$. The right figure is for tuning β with $\alpha = 2^{-7}$

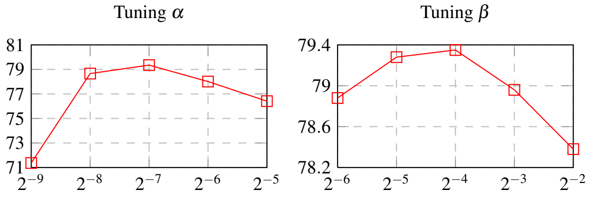

Figure 4. Parameter tuning on CMU PIE dataset with polynomial kernel. The left figure is for tuning α with $\beta = 2^{-4}$. The right figure is for tuning β with $\alpha = 2^{-7}$

CMU PIE dataset

The CMU PIE dataset contains 41,368 images of 68 individuals in total. Each individual is under 13 different poses, 43 different illumination conditions, and with 4 different expressions. Each individual thus may lie on multiple manifolds. Five near frontal poses ($C05$, $C07$, $C09$, $C27$, $C29$) and all different illuminations and expressions are used in our experiment. There are about 170 images for each individual and 11,554 images in total. The parameter tuning for α and β is reported in Figure 2,3 and 4 with linear kernel, Hellinger kernel and polynomial kernel, respectively. From Figure 2,3 and 4, the optimal α is 2^{-9}, 2^{-7} and 2^{-7} for linear kernel, Hellinger kernel and polynomial kernel, respectively. The optimal β is 2^{-4} for all three types of kernel. Table 1 shows the recognition rate of NN, SVM, CRC, SRC, CSDL, and KCSDL-LDA. From Table 1, our proposed KCSDL-LDA algorithm achieves superior performance over the other four classical classification methods and outperforms CSDL 5.09%, 6.66% and 6.18% for linear kernel, Hellinger kernel and polynomial kernel,respectively.

Extended YaleB dataset

For the Extended YaleB dataset, there are 2,414 frontal face images of 38 individuals in total. All the images are captured under varying illumination conditions. Similar to the parameters tuned on CMU PIE dataset, the optimal α is 2^{-9}, 2^{-7} and 2^{-7} for linear kernel, Hellinger kernel and polynomial kernel, respectively. The optimal β is 2^{-5}, 2^{-5} and 2^{-8} for linear kernel, Hellinger kernel and polynomial kernel, respectively. Table 2 shows the recognition rate of NN, SVM, CRC, SRC, CSDL, and KCSDL-LDA. From Table 2, our proposed KCSDL-LDA algorithm achieves superior performance to other four classical classification methods and outperforms CSDL 0.45%, 1.95% and 1.21% for linear kernel, Hellinger kernel and polynomial kernel,respectively.

Methods\kernels	linear	Hellinger	poly
NN	30.79 ± 1.73	NA	NA
SVM	65.35 ± 2.70	64.99 ± 2.55	65.38 ± 2.73
CRC	73.24 ± 2.33	75.24 ± 1.84	73.22 ± 2.27
SRC	72.24 ± 2.12	70.69 ± 2.14	69.15 ± 2.14
CSDL	74.66 ± 2.18	74.66 ± 1.97	73.78 ± 2.02
KCSDL-LDA	**79.75 ± 1.46**	**81.32 ± 1.37**	**79.96 ± 1.33**

Table 1: Recognition rate on the CMU PIE dataset (%).

Methods\kernels	linear	Hellinger	poly
NN	36.21 ± 2.51	NA	NA
SVM	65.50 ± 2.64	79.13 ± 2.68	65.26 ± 2.75
CRC	77.79 ± 2.23	88.60 ± 1.34	76.00 ± 2.30
SRC	77.61 ± 1.75	88.76 ± 1.86	75.55 ± 2.65
CSDL	78.42 ± 1.79	90.13 ± 1.49	78.37 ± 2.19
KCSDL-LDA	**78.87 ± 1.91**	**92.08 ± 1.68**	**79.58 ± 2.13**

Table 2: Recognition rate on the Extended YaleB dataset (%).

AR dataset

For the AR dataset, there are over 4,000 frontal faces for 126 individuals. A subset consisting of 50 male and 50 female categories is used. There are 26 face images for each class. Compared with the two above datasets, the AR dataset contains more facial variations, such as illumination change, various expressions, and facial disguises. Similar to the parameter tuning on CMU PIE dataset, the optimal α is 2^{-10}, 2^{-7} and 2^{-7} for linear kernel, Hellinger kernel and polynomial kernel, respectively. The optimal β is 2^{-4} for all three types of kernel. Table 3 shows the recognition rate of NN, SVM, CRC, SRC, CSDL, and KCSDL-LDA. From Table 3, our proposed KCSDL-LDA algorithm achieves superior performance over the other four classical classification methods and outperforms CSDL 2.68%, 3.85% and 3.86% for linear kernel, Hellinger kernel and polynomial kernel, respectively.

Methods\kernels	linear	Hellinger	poly
NN	30.50 ± 1.81	NA	NA
SVM	80.55 ± 1.11	80.45 ± 1.24	80.25 ± 1.04
CRC	91.68 ± 0.55	92.09 ± 0.53	92.00 ± 0.74
SRC	89.14 ± 1.07	85.86 ± 1.18	85.61 ± 1.05
CSDL	91.16 ± 1.04	89.49 ± 1.15	89.49 ± 1.08
KCSDL-LDA	**93.84 ± 0.92**	**93.34 ± 0.69**	**93.35 ± 1.08**

Table 3: Recognition rate on the AR dataset (%).

Conclusion

In this paper, we mainly focus on improving conventional class specific dictionary learning for face recognition. On one

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-456

hand, we extend the conventional class specific dictionary learning to arbitrary kernel space to capture nonlinear structures hidden in face images. On the other hand, we embed the linear discriminant analysis information into the class specific dictionary learning algorithm. These enhancements extremely improve the performance of face recognition rate. Experimental results demonstrate our proposed KCSDL-LDA algorithm for face recognition tasks.

Acknowledgment

This paper is supported partly by the National Natural Science Foundation of China (Grant No. 61402535, No. 61271407), the Natural Science Foundation for Youths of Shandong Province, China (Grant No. ZR2014FQ001) , Qingdao Science and Technology Project (No. 14-2-4-111-jch), and the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 14CX02169A).

References

[1] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, pages 1643–1650. IEEE, 2009.

[2] Yu-Xiong Wang and Yu-Jin Zhang. Image inpainting via weighted sparse non-negative matrix factorization. In *ICIP*, pages 3409–3412. IEEE, 2011.

[3] Bao-Di Liu, Yu-Xiong Wang, Yu-Jin Zhang, and Bin Shen. Learning dictionary on manifolds for image classification. *Pattern Recognition*, 46(7):1879–1890, 2013.

[4] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 31(2):210–227, 2009.

[5] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing*, 15(12):3736–3745, 2006.

[6] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

[7] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[8] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[9] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE Trans. PAMI*, 27(3):328–340, 2005.

[10] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, 1997.

[11] O Duda Richard, E Hart Peter, and G Stork David. Pattern classification. *A Wiley-Interscience*, pages 373–378, 2001.

[12] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. Geometric mean for subspace selection. *IEEE Trans. PAMI*, 31(2):260–274, 2009.

[13] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *ICCV*, pages 471–478. IEEE, 2011.

[14] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2009.

[15] Bao-Di Liu, Yu-Xiong Wang, Yu-Jin Zhang, and Yin Zheng. Discriminant sparse coding for image classification. In *ICASSP*, pages 2193–2196. IEEE, 2012.

[16] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691–2698. IEEE, 2010.

[17] Meng Yang, Lei Zhang, Jian Yang, and Dejing Zhang. Metaface learning for sparse representation based face recognition. In *ICIP*, pages 1601–1604. IEEE, 2010.

[18] Haoran Wang, Chunfeng Yuan, Weiming Hu, and Changyin Sun. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recogntiion*, 45(11):3902–3911, 2012.

[19] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *IJCV*, 109(3):209–232, 2014.

[20] Bao-Di Liu, Bin Shen, and Wang Yu-Xiong. Class specific dictionary learning for face recognition. In *ICSPAC*, pages 229–234, 2014.

[21] Bao-Di Liu, Liangke Gui, Yuting Wang, Yu-Xiong Wang, Bin Shen, Xue Li, and Yan-Jiang Wang. Class specific centralized dictionary learning for face recognition. *Multimedia Tools and Applications*, pages 1–19, 2015.

[22] Ning Zhou, Yi Shen, Jinye Peng, and Jianping Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, pages 3490–3497. IEEE, 2012.

[23] Shenghua Gao, Ivor Wai-Hung Tsang, and Yi Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *IEEE Trans. Image Processing*, 23(2):623–634, 2014.

[24] Bao-Di Liu, Bin Shen, and Xue Li. Locality sensitive dictionary learning for image classification. In *ICIP*, pages 3807–3811. IEEE, 2015.

[25] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *ICANN*, pages 583–588. Springer, 1997.

[26] Jianxin Wu and James M Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV*, pages 630–637. IEEE, 2009.

[27] Jan C van Gemert, Cor J Veenman, Arnold WM Smeulders, and J-M Geusebroek. Visual word ambiguity. *IEEE Trans. PAMI*, 32(7):1271–1283, 2010.

[28] Bao-Di Liu, Yu-Xiong Wang, Bin Shen, Yu-Jin Zhang, and Martial Hebert. Self-explanatory sparse representation for image classification. In *ECCV*, pages 600–616. Springer, 2014.

[29] Bao-Di Liu, Yu-Xiong Wang, Bin Shen, Xue Li, Yu-Jin Zhang, and Yan-Jiang Wang. Blockwise coordinate descent schemes for efficient and effective dictionary learning. *Neurocomputing*, 178:25–35, 2015.

[30] Bao-Di Liu, Yu-Xiong Wang, Shen Bin, Yu-Jin Zhang, and Yan-Jiang Wang. Blockwise coordinate descent schemes for sparse representation. In *ICASSP*, 2014.

[31] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI*, 23(6):643–660, 2001.

[32] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.

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

[34] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Author Biography

Bao-Di Liu received his BS in Signal and Information Processing from China University of Petroleum (2007) and his PhD in Electronic Engineering from Tsinghua University (2013). Currently, he is an assistant professor in College of Information and Control Engineering, China University of Petroleum, China. His research interests include computer vision and machine learning.