A Novel Multiple Kernel Sparse Representation based Classification for Face Recognition

Hao Zheng[1], Qiaolin Ye[2] and Zhong Jin[3]
[1] School of Mathematics and Information Technology, Nanjing XiaoZhuang University
Nanjing, China
[e-mail: gnjzhh@gmail.com]
[2] Computer science department, Nanjing Forestry University
Nanjing, China
[e-mail: yeqiaolincom@163.com]
[3] School of Computer Science and Technology, Nanjing University of Science and Technology
Nanjing, China
[e-mail: zhongjin@njust.edu.cn]

*Corresponding author: Hao Zheng

Received December 20, 2013; revised March 1, 2014; accepted March 25, 2014; published April 29, 2014

Abstract

It is well known that sparse code is effective for feature extraction of face recognition, especially sparse mode can be learned in the kernel space, and obtain better performance. Some recent algorithms made use of single kernel in the sparse mode, but this didn't make full use of the kernel information. The key issue is how to select the suitable kernel weights, and combine the selected kernels. In this paper, we propose a novel multiple kernel sparse representation based classification for face recognition (MKSRC), which performs sparse code and dictionary learning in the multiple kernel space. Initially, several possible kernels are combined and the sparse coefficient is computed, then the kernel weights can be obtained by the sparse coefficient. Finally convergence makes the kernel weights optimal. The experiments results show that our algorithm outperforms other state-of-the-art algorithms and demonstrate the promising performance of the proposed algorithms.

Keywords: sparse representation, dictionary, multiple kernels, face recognition

A preliminary version of this paper appeared in Springer CCPR 2012, September 24-26, Beijing, China. This version includes detailed introduction, related work, illustration of convergence of algorithm, kernel parameters selection and experiment results on more databases.This research was supported the National Science Foundation of China under grant No. 60973098.

http://dx.doi.org/10.3837/tiis.2014.04.017

1. Introduction

Over the last decade, there has been rapid development in Image recognition, and many algorithms have been proposed, such as Eigenface[1] , locality preserving projection(LPP) [2] , Fisherface [3], maximum margin criterion (MMC) [4]. In a recent work, Yu and Tao[5] proposed an adaptive hypergraph learning method for transductive image classification. Afterward, a semi-supervised classification algorithm named semi-supervised multiview distance metric learning[6] were proposed. To efficiently combine the visual features for subsequent retrieval and synthesis tasks , another new semi-multiview subspace learning algorithm[7] was proposed. In addition, Wang et al. [8] proposed a neighborhood similarity measure to explores the local sample and label distributions. To integrate multiple complementary graphs into a regularization framework, the optimized multigraph-based semi-supervised learning algorithm[9] was subsequently proposed. In addition other related methods[10-14] were proposed. In Image recognition, Face recognition is a very challenging question, especially classifier is important for its final role. The nearest-neighbor(NN) algorithm is extremely simple and it is accurate and applicable to various problems [15]. The simplest 1-nn algorithm assigns an input sample to the category of its nearest neighbor from the labeled training set. Due to the NN's shortcoming that only one training sample is used to represent the test face image, the nearest feature line classifier [16] was proposed through using two training samples for each class to represent the test face image. Then the nearest feature plane classifier [17] was proposed through using three samples to represent the test image. Later, for representing the test image by all the training samples of each class, the local subspace classifier [18] and the nearest subspace classifier [19-21] were proposed. The support vector machine (SVM) classifier is also another classifier which is solidly based on the theory of structural risk minimization in statistical learning. It is well known that the SVM maps the inputs to a high-dimensional feature space and then finds a large margin hyperplane between the two classes which can be solved through the quadratic programming algorithm. For noise images, sparse representation based classifier (SRC) [22] make good performance, and SRC shows exciting results in dealing with occlusion by assuming a sparse coding residual. Later, many extended algorithms were proposed, e.g. Gabor SRC [23], SRC for face misalignment or pose variation [24-25], SRC for continuous occlusion [26] and Heteroscedastic SRC [27].

Recently the kernel approach [28] has attracted great attention. It offers an alternative solution to increase the computational power of linear learning machines by mapping the data into a high dimensional feature space. The approach has been studied and extended some kernel based algorithms such as kernel principal component analysis (KPCA) [29] and kernel fisher discriminant analysis (KFD)[30, 31]. As the extension of conventional nearest-neighbor algorithm, the kernel optimization algorithm [32-38] was proposed which can be realized by substitution of a kernel distance metric for the original one in Hilbert space. By choosing an appropriate kernel function, the results of kernel nearest-neighbor algorithm are better than those of conventional nearest-neighbor algorithm. Similarity, the single-kernel SVM classifier was proposed, and various remedies were introduced, such as the reduced set method [39],[40],bottom-up method[41], building of a sparse large margin classifier[42],[43], and the incremental building of a reduced-complexity classifier.

But above methods have some disadvantages. NN predicts the category of the image to be tested by only using its nearest neighbor in the training data, and it can easily be affected by

noise. NS approximates the test image to the category which minimizes the reconstruction error, therefore the performance is not ideal when the classes are highly correlated to each other. The shortcoming of the SVM is that it is often not as compact as the other classifiers such as neural networks. Fortunately Wright et al. [22] proposed a sparse representation based classifier for face recognition (SRC) which first codes a testing sample as a sparse linear combination of all the training samples, and then classifies the testing sample by evaluating which class leads to the minimum representation error. SRC is much more effective than state-of-art methods in dealing with face occlusion, corruption, lighting and expression changes, etc. It is well known that if an appropriate kernel function is utilized for a test sample, more neighbors probably have the same class label in the high dimensional feature space. Sparse representation in the high dimensional space can improve the performance of recognition and discriminative ability. Some methods were proposed such as kernel representation based classification algorithm (KSRC) [44-46], etc. However the algorithm is often unclear about what is the most suitable kernel for the task at hand, and hence the user may wish to combine several possible kernels. One problem with simply adding kernels is that using uniform weights is possibly not optimal. To overcome it, we proposed a novel algorithm named multiple kernel sparse representation based classifier (MKSRC) which can optimize the kernel weights while training the dictionary. The contributions of this paper can be summarized as follow.

1) We propose a multiple kernel sparse representation based classifier. By making full use of the kernel information, classification performance is improved compared with the state-of-the-art classifier.
2) Through the dicaitonary learning, kernel weights can be adaptive selected. Due to automaticly adjusting the weights, our classifier is more robust, especially for occlusion images.
3) We conduct the experiments in two facial image databases in the conditions of no occlusion and block occlusion. The experiments results validate the effectiveness of new classifier.

2. Related Work

2.1 Sparse representation based classification

Sparse representation based classification (SRC) was reported by Wright [22] for robust face recognition. In Wright 's pioneer work, the training face images are used as the dictionary of representative samples, and an input test image is coded as a sparse linear combination of these sample images via l_1-norm minimization.

Given a signal (or an image) $y \in \Re^m$, and a matrix $A = [a_1, a_2, \cdots, a_n] \in \Re^{m \times n}$ containing the elements of an overcomplete dictionary in its columns, the goal of sparse representation is to represent y using as few entries of A as possible. This can be formally expressed as follows:

$$\hat{x} = \arg\min \| x \|_0, \ s.t. \ y = Ax \tag{1}$$

where $x \in \Re^n$ is the coefficient vector, and $\| x \|_0$ is the l_0-norm which is equal to the number of non-zero components in x. However, this criterion is not convex, and finding the sparsest solution of Eq. (1) is NP-hard. Fortunately this difficulty can be overcomed by convexizing the

problem and solving

$$\hat{x} = \arg\min \| x \|_1 , \ s.t. \ y = Ax \tag{2}$$

where l_1 is used instead of l_0. It can be shown that if the solution x sought is sparse enough, the solution of l_0 minimization problem is equal to the solution of l_1 minimization problem.

Finally, for each class i, let $\delta_i : \Re^n \to \Re^n$ be the characteristic function which selects the coefficients associated with the i-th class. Using only the coefficients associated with the i-th class, one can approximately reconstruct the test sample y as $\tilde{y} = A\delta_i(\hat{x})$, then classify y based on these approximations by assigning it to the class that minimizes the residual:

$$r_i(y) = \| y - A\delta_i(\hat{x}) \|_2 , \ for \ i = 1,...,k. \tag{3}$$

If $r_l(y) = \min r_i(y)$, y is assigned to class l.

Now suppose that the face image is partially occluded or corrupted, the problem can be expressed as follows:

$$\hat{x} = \arg\min \| x \|_1 , \ s.t. \ y = Ax + \varepsilon \tag{4}$$

where ε is residual. We can approximately reconstruct the test sample y as $\tilde{y} = A\delta_i(\hat{x}) + \hat{\varepsilon}$, then compute the residuals:

$$r_i(y) = \| y - A\delta_i(\hat{x}) - \hat{\varepsilon} \|_2 , \ for \ i = 1,...,k. \tag{5}$$

If $r_l(y) = \min r_i(y)$, y is assigned to class l.

2.2. Kernel sparse representation based classification (KSRC) [44]

It is well known that kernel approach can change the distribution of samples through mapping samples into a high dimensional feature space by a nonlinear mapping. In the high dimensional feature space, the sample can be represented more accurately by sparse representation dictionary.

Suppose there are p classes in all, and the set of the training samples is $A = [A_1, A_2, \ldots, A_p] = [x_{1,1}, x_{1,2}, \ldots, x_{p,n_p}] \in \Re^{d \times N}$, and $N = \sum_{i=1}^{p} n_i$ is total training samples number, $y \in \Re^{d \times 1}$ is test sample. The samples are mapped from original feature space into a high dimensional feature space : $y \to \varphi(y), A = [x_{1,1}, x_{1,2}, \ldots, x_{p,n_p}] \to U = [\varphi(x_{1,1}), \varphi(x_{1,2}), \cdots, \varphi(x_{p,n_p})]$ by a nonlinear mapping $\varphi : \Re^d \to \Re^k (d < K)$. The sparse representation mode can be formulated as

$$\hat{v} = \arg\min_{s} \left\| v \right\|_0 , \ s.t. \ \phi(y) = Uv \tag{6}$$

where $\phi(y)$ is test sample in the high dimensional feature space. Due to NP hard problem, the solution of Eq.(6) can be obtained through the following Eq.(7):

$$\hat{v} = \arg\min_{s} \|v\|_1, \ \ s.t. \ \ \phi(y) = Uv. \tag{7}$$

In the presence of noises, the Eq.(7) should be relaxed and the following optimization problem is obtained:

$$\hat{v} = \arg\min_{s} \|v\|_1, \ \ s.t. \ \ \|Uv - \phi(y)\|_2 \le \varepsilon. \tag{8}$$

Though U and $\phi(y)$ are unknown, according to [44],we can prove that Eq.(8) is equivalent to the following Eq.(9):

$$\hat{v} = \arg\min_{v} \|v\|_1, \ \ s.t. \ \ \|U^T Uv - U^T \phi(y)\|_2 \le \delta. \tag{9}$$

where

$$U^T U = [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{c,l_c})]^T [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{p,n_p})]$$

$$= \begin{bmatrix} K(x_{1,1}, x_{1,1}) & K(x_{1,1}, x_{1,2}) & \cdots & K(x_{1,1}, x_{p,n_p}) \\ K(x_{1,2}, x_{1,1}) & K(x_{1,2}, x_{1,2}) & \cdots & K(x_{1,2}, x_{p,n_p}) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_{p,n_p}, x_{1,1}) & K(x_{p,n_p}, x_{1,2}) & \cdots & K(x_{p,n_p}, x_{p,n_p}) \end{bmatrix}. \tag{10}$$

$$U^T \phi(y) = [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{p,n_p})]^T \phi(y)$$

$$= \begin{bmatrix} K(x_{1,1}, y) \\ K(x_{1,2}, y) \\ \vdots \\ K(x_{p,n_p}, y) \end{bmatrix}. \tag{11}$$

The procedures of KSRC algorithm are summarized as Algorithm1:

Algorithm 1 (kernel sparse representation based classification)

Input:

the training samples matrix

$A=[A_1,A_2,\ldots,A_p]=[x_{1,1},x_{1,2},\ldots,x_{p,n_p}]\in\Re^{d\times N}$, $N=\sum_{i=1}^{p}n_i$, test sample $y\in\Re^d$, $\varepsilon>0$, a kernel function.

output: identity(y).

Step 1: normalize the matrix A.

Step 2: obtain U^TU 和$U^T\phi(y)$ by Eq. (10) and Eq. (11).

Step 3: obtain $\hat{v}$ by solving the l_1-norm problem by Eq. (9).

Step 4: compute the residuals: $r_i(y)=\|U^T\phi(y)-U^TU\hat{v}\|_2$.

Step 5: compute identity(y) = $\arg\min_k(r_k(y))$.

3. Multiple Kernel Sparse Representation based Classifier (MKSRC)

Suppose there are p classes in all, and the set of the training samples is $A=[A_1,A_2,\ldots,A_p]=[x_{1,1},x_{1,2},\ldots,x_{p,n_p}]\in\Re^{d\times\sum_{i=1}^{p}n_i}$, and $y\in\Re^{d\times1}$ is the test sample. The traditional sparse coding model is equivalent to the so-called LASSO problem [47]:

$$\min\|y-A\alpha\|_2^2,\ s.t.\|\alpha\|_1<\sigma. \qquad \text{where } \sigma>0 \text{ is a constant.}$$

Suppose there is a feature mapping function $\varphi:\Re^d\to\Re^k(d<K)$. It maps the feature and basis to the high dimensional feature space:

$y\to\varphi(y), A=[x_{1,1},x_{1,2},\ldots,x_{p,n_p}]\to U=[\varphi(x_{1,1}),\varphi(x_{1,2}),\cdots,\varphi(x_{p,n_p})]$ .There exits one problem that one kernel is not most suitable kernel, so we wish to combine several possible kernels. Multiple kernel sparse representation based classification (MKSRC) is a way of optimizing kernel weights while training dictionary. The mode of Multiple Kernel by Lanckriet [48] is $k(x_i,x_j)=\sum_{k=1}^{m}\alpha_k k_k(x_i,x_j)$, and we restrain the kernel weights by $\sum_{i=1}^{m}{\alpha_K}^2=1,\ \alpha_K\geq0$, then substitute the mapped features and basis to the formulation of sparse coding, obtain the objective function as follows:

$$\min\|\phi(y)-Uv\|_2^2,\ s.t.\|v\|_1<\sigma\ \sum_{i=1}^{m}{\alpha_K}^2=1. \qquad (12)$$

Table 1. Important notations used in this paper and their description

Notation	Description
φ	Mapping function from original dimensional space to high dimensional space
A	The matrix of training samples in original space
U	The matrix of training samples in high dimensional space
α_k	Kernel weights
$k_k(x_i, x_j)$	Kernel function
v	The vector for sparse coefficient in high dimensional space
p	The number of classes in the database
m	The number of kernel function

The Lagrangian function for Eq. (12) is :

$$\begin{aligned} J &= \| \phi(y) - Uv \|_2^2 + \lambda \| v \|_1 + \gamma(\boldsymbol{\alpha}^T \boldsymbol{\alpha} - 1) \\ &= \phi(y)^T \phi(y) - 2v^T U^T \phi(y) + v^T U^T U v + \lambda \| v \|_1 + \gamma(\boldsymbol{\alpha}^T \boldsymbol{\alpha} - 1) \\ &= \sum_{k=1}^{m} \alpha_k k_k(y, y) - 2v^T [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{p,n_p})]^T \phi(y) \\ &\quad + v^T [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{p,n_p})]^T [\phi(x_{1,1}), \phi(x_{1,2}), \cdots, \phi(x_{p,n_p})] v + \lambda \| v \|_1 + \gamma(\boldsymbol{\alpha}^T \boldsymbol{\alpha} - 1). \end{aligned} \tag{13}$$

For sample x and y , we have:

$$\phi(x_i)^T \phi(y_j) = k(x_i, y_j) \;\; k(x_i, x_j) = \sum_{k=1}^{m} \alpha_k k_k(x_i, x_j) \; .$$

Therefore

$$\begin{aligned} J = \sum_{k=1}^{m} \alpha_k k_k(y, y) - 2v^T & \begin{bmatrix} \sum_{k=1}^{m} \alpha_k k_k(x_{1,1}, y) \\ \sum_{k=1}^{m} \alpha_k k_k(x_{1,2}, y) \\ \vdots \\ \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p}, y) \end{bmatrix} + v^T \begin{bmatrix} \sum_{k=1}^{m} \alpha_k k_k(x_{1,1}, x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{1,1}, x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{1,1}, x_{p,n_p}) \\ \sum_{k=1}^{m} \alpha_k k_k(x_{1,2}, x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{1,2}, x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{1,2}, x_{p,n_p}) \\ \vdots & \vdots & \ddots & \\ \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p}, x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p}, x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p}, x_{p,n_p}) \end{bmatrix} v \\ & + \lambda \| v \|_1 + \gamma(\sum_{k=1}^{m} \alpha_k{}^2 - 1). \end{aligned} \tag{14}$$

Setting the derivative of J w.r.t. the primal variable α_k to zero,

$$\frac{\partial J}{\partial \alpha_K} = k_K(y,y) - 2v^T \begin{bmatrix} k_k(x_{1,1},y) \\ k_k(x_{1,1},y) \\ \vdots \\ k_k(x_{p,n_p},y) \end{bmatrix} + v^T \begin{bmatrix} k_k(x_{1,1},x_{1,1})\ k_k(x_{1,1},x_{1,2}) & \cdots & k_k(x_{1,1},x_{p,n_p}) \\ k_k(x_{1,2},x_{1,1})\ k_k(x_{1,2},x_{1,2}) & \cdots & k_k(x_{1,2},x_{p,n_p}) \\ \vdots & \vdots & \ddots \\ k_k(x_{p,n_p},x_{1,1}) k_k(x_{p,n_p},x_{1,2}) & \cdots & k_k(x_{p,n_p},x_{p,n_p}) \end{bmatrix} v + 2\gamma\alpha_K = 0. \tag{15}$$

Finally we obtain:

$$\alpha_K = -\frac{1}{2\gamma}(k_i(y,y) - 2v^T \begin{bmatrix} k_k(x_{1,1},y) \\ k_k(x_{1,1},y) \\ \vdots \\ k_k(x_{p,n_p},y) \end{bmatrix} + v^T \begin{bmatrix} k_k(x_{1,1},x_{1,1})\ k_k(x_{1,1},x_{1,2}) & \cdots & k_k(x_{1,1},x_{p,n_p}) \\ k_k(x_{1,2},x_{1,1})\ k_k(x_{1,2},x_{1,2}) & \cdots & k_k(x_{1,2},x_{p,n_p}) \\ \vdots & \vdots & \ddots \\ k_k(x_{p,n_p},x_{1,1}) k_k(x_{p,n_p},x_{1,2}) & \cdots & k_k(x_{p,n_p},x_{p,n_p}) \end{bmatrix} v). \tag{16}$$

Because $\phi(y)$ and U are unknown, Eq. (12) cannot be solved directly. But according to [34], Eq. (12) can be transformed to

$$\hat{v} = \arg\min\{\| U^T\phi(y) - U^T U v \|_2^{\ 2} + \lambda \| v \|_1\}. \tag{17}$$

where

$$U^T\phi(y) = \begin{bmatrix} \sum_{k=1}^{m} \alpha_k k_k(x_{1,1},y) \\ \sum_{k=1}^{m} \alpha_k k_k(x_{1,2},y) \\ \vdots \\ \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p},y) \end{bmatrix}. \tag{18}$$

$$U^T U = \begin{bmatrix} \sum_{k=1}^{m} \alpha_k k_k(x_{1,1},x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{1,1},x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{1,1},x_{p,n_p}) \\ \sum_{k=1}^{m} \alpha_k k_k(x_{1,2},x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{1,2},x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{1,2},x_{p,n_p}) \\ \vdots & \vdots & \ddots & \\ \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p},x_{1,1}) & \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p},x_{1,2}) & \cdots & \sum_{k=1}^{m} \alpha_k k_k(x_{p,n_p},x_{p,n_p}) \end{bmatrix}. \tag{19}$$

Since initial weights are an estimator which is not optimal, the implementation of MKSRC is an iterative process. When the difference of weights α_i is small enough, the convergence is

stopped. It can be formulated as follows: $|| \alpha^{t+1} - \alpha^{t} || \leq tol$. In order to verify the convergence of the MKSRC algorithm, Experiments on ORL database were done. It is straightforward that the proposed MKSRC algorithm converges because recognition rate is stable after several iterations, as illustrated in **Fig. 1**

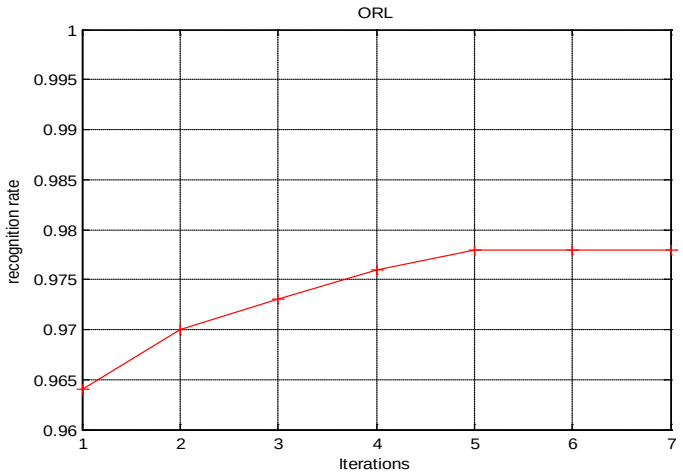

Fig. 1. Illustration of the convergence of Algorithm 2

The MKSRC algorithmic procedures can be summarized as Algorithm 2:

Algorithm 2 (Multiple kernel sparse representation based classification)

Step 1: Input training samples $A \in \Re^{d \times \sum_{i=1}^{p} n_i}$ partitioned into p classes, and a test sample y , the number of kernel function m .

Step2: Compute initial weights $\alpha_K = \frac{1}{\sqrt{m}}$ and $k(x_i, x_j) = \sum_{k=1}^{m} \alpha_k k_k(x_i, x_j)$.

Step 3: Compute the coefficient $\hat{v} = \arg\min \{ || U^T \phi(y) - U^T U v ||_2^{\ 2} + \lambda || v ||_1 \}$ by Eq. (18) and Eq. (19).

Step 4: Compute the weights α_K by Eq. (16).

Step 5: Go back to step 3 until the condition of convergence is met.

Step 6: Compute $r_j(y) = || U^T \phi(y) - U^T U \hat{v} ||_2^{\ 2}$.

Step 7: Output that $\text{identity}(y) = \arg\min r_j(y)$.

4. Experiments and discussions

In this section, we perform experiments on face databases to demonstrate the efficiency of MKSRC. To evaluate more comprehensively the performance of MKSRC, in section 4.1 we discuss the comparison methods and experiment configurations, then in section 4.2 test FR without occlusion, and finally in section 4.3 we test FR with block occlusion. Through experiments we chose three kernel functions: linear kernel, polynomial kernel, and gussian kernel, of which the kernel parameters were tuned using cross validation. For statistical stability, we generate ten different training and test dataset pairs by randomly permuting 10

times. We compare the performance of the proposed MKSRC with the state-of-the-art classifiers, such as SVM [41], SRC [22], KSRC (Polynomial) [44], KSRC (Gaussian) [44].

4.1 Comparison methods and configurations

To verify the performance of the MKSRC method, we selected the following the methods to compare.

1) SVM. Here, we use the one-versus-all strategy, and select RBF kernel. The radius parameter is tuned to their optimal values through cross validation.
2) SRC. SRC is an effective classifier which codes a testing sample as a sparse linear combination of all the training samples. In order to obtain the better performance, we select the basic pursuit algorithm. The parameter value 0.001 of the SRC is selected by cross validation.
3) KSRC (Polynomial). KSRC makes use of polynomial kernel function to improve the classifier performance. Through the experiments without occlusion, the value of kernel parameter in FERET face database is set to 2, the value in ORL face database is set to 13, while in block occlusion condition both of the values are set to 2.
4) KSRC (Gaussian). KSRC makes use of Gaussian kernel function to improve the classifier performance. Through the experiments without occlusion, both of the values in FERET and ORL face database are set to 2, while in block occlusion condition, the value of kernel parameter in FERET face database is set to 2, the value in ORL face database is set to 3.

4.2 Face recognition without occlusion

1) The FERET face dataset

FERET database [49] were used in our experiments including the images marked with two-character strings, i.e., “ba,” “bj,” “be,” “bk” “bf,” “bd,” and “bg.” Thus, the entire data set include 1400 images of 200 different subjects, with 7 images per subject. All these images were aligned according to the center points of the eyes. The images are of size 80 by 80. Some sample images are shown in **Fig. 2**. The 800 images of 200 subjects were randomly used for training, while the remaining 600 of 200 subjects were used for testing. **Table 1** and **Fig. 3** show the recognition rate in different algorithms. We can see that MKSRC algorithm retains higher performance than SRC and KSRC. In dimension 300, the recognition rate of MKSRC is 69.32% which is 4.5% higher than SVM.

Fig. 2. Sample images of one person on FERET face database

For the selection of kernel parameters, we find the candidate interval from 1 to 10. For simple computing, we find the optimal kernel parameters within these intervals through the single kernel experiments. **Fig. 4**(a) shows the recognition rates of the polynomial kernel versus the variation of the parameter d, **Fig. 4**(b) shows the recognition rates of the gaussian kernel versus the variation of the parameter t. From the **Fig.** 4, both of optimal parameter t and d are 2.

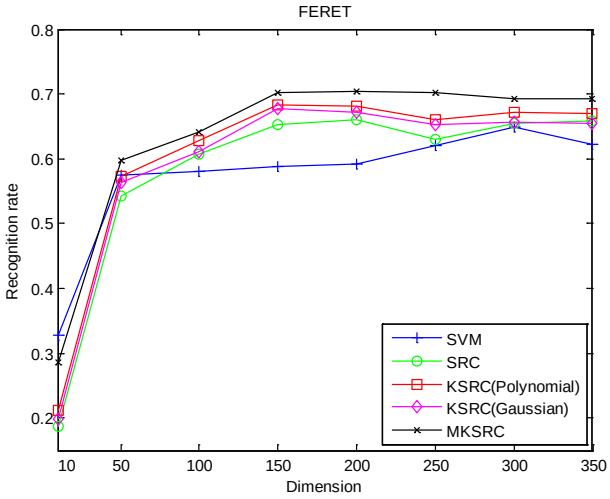

Fig. 3. The average recognition rates of SVM, SRC, KSRC (Polynomial), KSRC (Gaussian) and MKSRC versus the dimensions on FERET face database

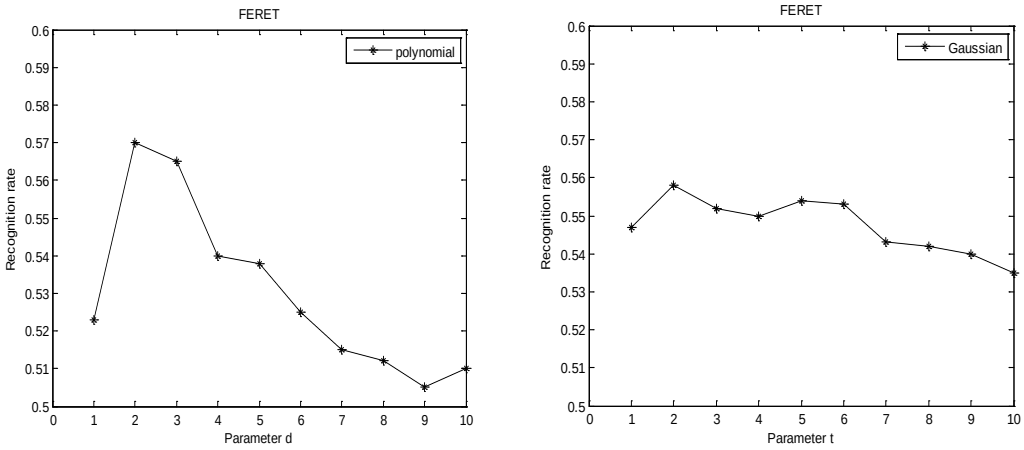

Fig. 4. (a) Recognition rates versus the parameter d of polynomial kernel (b) Recognition rates versus the parameter t of Gaussian kernel

Table 1. Accuracy on FERET face database

	SVM	SRC	KSRC(Polynomial)	KSRC(Gaussian)	MKSRC
Dimensions(d=50)					
Accuracy	57.5%	54.3%	57.3%	56.4%	59.8%
Parameter			d=2	t=2	d=2 t=2
Dimensions(d=200)					
Accuracy	59.2%	66%	68.2%	67.1%	70.3%
Parameter			d=2	t=2	d=2 t=2
Dimensions(d=300)					
Accuracy	64.8%	65.5%	67.1%	65.6%	69.32%
Parameter			d=2	t=2	d=2 t=2

2) The ORL face dataset

The ORL face database consists of 400 frontal face images of 40 subjects. They are captured under various lighting conditions and cropped and normalized to 112 × 92 pixels. The face images were captured under various illumination conditions. We randomly split the database into two halves. One half (5 images per person) was used for training, and the other half for testing. The images are reduced to 30, 60, 110 dimensions, respectively. **Table 2** and **Fig. 5** illustrate the face recognition rates by different methods. We can see that the recognition rates increase with the larger dimensions. Our MKSRC algorithm achieves a recognition rate between 89% and 97.8%, much better than the other algorithms, especially in dimension 60 MKSRC gets the best performance.

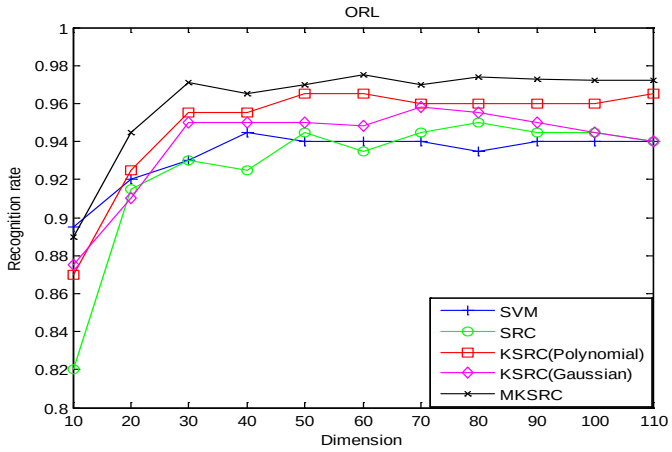

Fig. 5. The average recognition rates of SVM, SRC, KSRC (Polynomial), KSRC (Gaussian) and MKSRC versus the dimensions on ORL face database

Table 2. Accuracy on ORL face database

	SVM	SRC	KSRC(Polynomial)	KSRC(Gaussian)	MKSRC
Dimensions(d=30)					
Accuracy	93%	93%	95.5%	95.5%	97.1%
Parameter			d=13	t=2	
Dimensions(d=60)					
Accuracy	94%	93.5%	96.5%	95.5%	97.8%
Parameter			d=13	t=2	
Dimensions(d=110)					
Accuracy	94%	94%	96.5%	94%	97.2%
Parameter			d=13	t=2	

From the experiment without occlusion, we can see the proposed MKSRC method not only outperforms the SRC, but also outperforms the KSRC. Experiments results demonstrate that kernel information helps to improve the recognition rate. This is attributed to two reasons: 1) face image features in kernel feature space contain more effective discriminant information than features in the original feature space, therefore the samples can be easily separated. 2) The appropriate kernel combination makes the test sample in the high dimensional feature

space reflect its class label information more accurately. In addition, different kernel functions conduct different experiments results, so the selection of the kernel functions and their kernel parameters is important.

4.3 Face recognition with block occlusion

1) FERET database

The next experiment is about occlusion for FERET database. We randomly take the four face images of each person for training and the rest three face images for testing. We simulate various levels of contiguous occlusion, from 10% to 30%, by replacing a randomly located square block of each test image with an unrelated image, Again, the location of occlusion are randomly chosen for each image and are unknown to the computer. For computer convenience, the dimension is reduced to 50.

Table 3. Accuracy on FERET face database under occlusion

	SRC	KSRC(Polynomial)	KSRC(Gaussian)	MKSRC
Occlusion(10%)				
Accuracy	38.2%	41.4%	40.3%	42.1%
Parameter		d=2	t=2	d=2 t=2
Occlusion (20%)				
Accuracy	31.3%	34.2%	33.5%	36.6%
Parameter		d=2	t=2	d=2 t=2
Occlusion (30%)				
Accuracy	25.5%	27.8%	27.1%	29.2%
Parameter		d=2	t=2	d=2 t=2

From **Table 3** we can see that the accuracy rate of all the methods decline with the occlusion levels increasing, which indicates that loss of feature affects the face recognition performance. But MKSRC achieves better performance than other algorithms. When occlusion is 30%, SRC is only 25.5%, while MKSRC is 29.2% which is more than 3.7% improvement than SRC.

2) ORL database with regular shapes occlusion

The next one is that we test the efficiency of MKSRC to the block occlusion using the ORL face dataset. We randomly take the first half for training and the rest for testing. We simulate various levels of contiguous occlusion, from 10% to 30%, by replacing a randomly located square block of each test image with an unrelated image, Again, the location of occlusion is randomly chosen for each image and is unknown to the computer. A test example of ORL with 30% occluded block is shown as **Fig. 6**. Here, for computational convenience, the size of image is cropped to 32 × 32. The dimensions of the images are reduced to 60. The results of the experiments are more exciting. From **Table 4** we can see that the accuracy rate of all the methods decline with the occlusion levels increasing, which indicates that loss of feature affects the face recognition performance. But MKSRC retains good performance of 74.8% when the occlusion percentage is 30%. Through above experiments the fact has been verified that combination of the multiple kernels can improve the performance of face recognition.

Fig. 6. A test example of ORL face database with 30% occluded block

Table 4. Accuracy on ORL face database under occlusion

	SRC	KSRC(Polynomial)	KSRC(Gaussian)	MKSRC
Occlusion(10%)				
Accuracy	89%	91%	90.4%	93.3%
Parameter		d=2	t=3	d=2 t=3
Occlusion (20%)				
Accuracy	80.5%	83.5%	81%	84%
Parameter		d=2	t=3	d=2 t=3
Occlusion (30%)				
Accuracy	71%	73.6%	71%	74.8%
Parameter		d=2	t=3	d=2 t=3

3) ORL database with irregular shapes occlusion

The next one is more challenge, and we chose the irregular shape occlusion such as conch. The location of occlusion is randomly chosen for each image and is unknown to the computer. A test example of ORL with irregular shape occluded block is shown as Fig. 7. Here, for computational convenience, the size of image is cropped to 32 × 32. The dimensions of the images are reduced to 60. From Table 5 we can see that the MKSRC method retains the good performance, and accuracy is 5.7% than SRC. This demonstrates that the MKSRC method is stable, and suitable in the different occlusion conditions.

Fig. 7. A test example of ORL face database with irregular shape occluded block

Table 5. Accuracy on ORL face database under irregular shape occlusion

	SRC	KSRC(Polynomial)	KSRC(Gaussian)	MKSRC
Accuracy	83.3%	86.7%	85%	89%
Parameter		d=2	t=3	d=2 t=3

The face exeriments with block occlusion demostrate the MKSRC method is more robust than other methods. We conduct exhaustive experiments not only in two face image database, but also in the conditions of regular shape occlusion and irregular shape occlusion. Because kernel weights can be adaptive selected, the MKSRC method get more suitable kernel combination, as a result achieve the better permance than other methods. With the occlusion rate increasing , the performance of the proposed method doesn't decline significantly. This means that the mutiple kernel classifier is not sensitive.

5. Conclusion

This paper proposed a multiple kernel sparse representation based classification. On the high-dimensional data such as face images, KSRC algorithm has got better performance than SRC, but KSRC algorithm does not make full use of kernel information. MKSR algorithm can solve this problem by combining several possible kernels, e.g. gussian kernel, while selecting the suitable weights of kernel function. On various face databases MKSRC algorithm achieves the best performance. Because kernel parameter is important for the recognition performance, we will focus on estimating the kernel parameter in the future.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. Article (CrossRef Link)
[2] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, 2005. Article (CrossRef Link)
[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 19, no. 7, pp. 711–720, 1997. Article (CrossRef Link)
[4] Li H, Jiang T, Zhang K., "Efficient and robust feature extraction by maximum margin criterion," In *Proc. of Proceedings of the advances in neural information processing systems*, vol.16. MIT Press, Vancouver, Canada, 2004. Article (CrossRef Link)
[5] J. Yu and D. Tao, "Adaptive Hypergraph Learning and Its Application in Image Classification," *IEEE Trans. on Image Processing*, vol. 21, no. 7, pp. 3262-3271, 2012. Article (CrossRef Link)
[6] J. Yu , M. Wang and D. Tao, "Semisupervised Multiview Distance Metric Learning for Cartoon Synthesis," *IEEE Trans. on Image Processing*, vol. 21, no. 11, pp. 4636-3271, 2012. Article (CrossRef Link)
[7] J. Yu , D. Liu and D. Tao, "On Combining Multiview Features for Cartoon Character Retrieval and Clip Synthesis," *IEEE Trans. on Systems*, Man, and Cybernetics, vol. 42, no. 5, pp. 1413-1427, 2012. Article (CrossRef Link)
[8] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, and Y. Song, "Unified Video Annotation via Multigraph Learning," *IEEE Trans. on Circuits and System for Video Technology*, vol. 19, no. 5, pp. 733-746, 2009. Article (CrossRef Link)
[9] M. Wang, X. Hua, J. Tang, and R. Hong, "Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation," *IEEE Trans. on Multimedia*, vol. 11, no. 3, pp. 465-476, 2009. Article (CrossRef Link)
[10] M. Wang, B. Ni, X. Hua, and T. Chua, "Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration," *ACM Computing Surveys*, vol. 4, no. 4, Article 25, 2012. Article (CrossRef Link)
[11] M. Wang, and X. Hua. "Active Learning in Multimedia Annotation and Retrieval: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2, pp. 10-31, 2011. Article (CrossRef Link)
[12] Y. Gao, M. Wang, Z. Zha, and J. Shen, "Visual Texttual Joint Relevance Learning for Tag-Based Social Image Search," *IEEE Trans. on Image Processing*, vol. 22, no. 1, pp. 363-376, 2012. Article (CrossRef Link)
[13] Y. Gao, M. Wang, D. Tao, and R. Ji, "3D Object Retrieval and Recognition with Hypergraph Analysis," *IEEE Trans. on Image Processing*, vol. 21, no. 9, pp. 4290-4303, 2012. Article (CrossRef Link)
[14] Y. Gao, M. Wang, Z. Zha, and Q. Tian, "Less is More: Efficient 3D Object Retrieval with Query View Selection," *IEEE Trans. on Multimedia*, vol. 13, no. 5, pp. 41007-1018, 2011. Article (CrossRef Link)
[15] Duda, R.O. and Hart, P.E., "Pattern Classification and Scene Analysis," *Wiley*, New York, 1973.

Article (CrossRef Link)
[16] S.Z. Li and J. Lu, "Face recognition using nearest feature line method," *IEEE Trans. Neural Network*, vol. 10, no. 2, pp. 439-443, 1999. Article (CrossRef Link)
[17] J.T. Chien, and C.C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644-1649, 2002. Article (CrossRef Link)
[18] J. Laaksonen, "Local subspace classifier", in *Proc. of Int'l Conf. Artificial Neural Networks*, 1997. Article (CrossRef Link)
[19] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005. Article (CrossRef Link)
[20] S.Z. Li, "Face recognition based on nearest linear combinations," in *Proc. of IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 1998. Article (CrossRef Link)
[21] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106-2112, 2010. Article (CrossRef Link)
[22] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., "Robust face recognition via sparse representation," *TPAMI*, vol. 31(2), pp. 210−227, 2009. Article (CrossRef Link)
[23] M. Yang and L. Zhang, "Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary," in *Proc. of European Conf. Computer Vision*, 2010. Article (CrossRef Link)
[24] J.Z. Huang, X.L. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, 2008. Article (CrossRef Link)
[25] A. Wagner, J. Wright, A. Ganesh, Z.H. Zhou, and Y. Ma, "Towards a Practical Face Recognition System: Robust Registration and Illumination by Sparse Representation," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, 2009. Article (CrossRef Link)
[26] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proc. of IEEE Int'l Conf. Computer Vision*, 2009. Article (CrossRef Link)
[27] H. Zheng, J. Xie, Z. Jin, "Heteroscedastic Sparse Representation Classification for Face Recognition," *Neural Processing Letters*, Vol. 35, Issues 3, pp 233-244, 2012. Article (CrossRef Link)
[28] Aizerman, M. A., Braverman, E. M. and Rozonoer, L. I., "Theoretical foundation of potential function method in pattern recognition learning," *Automat. Remote Contr.25*, 821-837, 1964. Article (CrossRef Link)
[29] B. Scholkopf, S. Alexander, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput. 10*, 1299-1319, 1998. Article (CrossRef Link)
[30] S. Mike, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," in *Proc. of Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, vol. IX, pp. 41–48, 1999. Article (CrossRef Link)
[31] S. Mike, G. Ratsch, B. Scholkopf, A. Smola, J. Weston, and K.R. Muller, "Invariant feature extraction and classification in kernel spaces," in *Proc. of Proceedings of the 13th Annual Neural Information Processing Systems Conference*, pp. 526–532, 1999. Article (CrossRef Link)
[32] A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil, "A DC algorithm for kernel selection," in *Proc. of 23rd Int. Conf. Mach.*, Pittsburgh,PA, pp. 41–49, 2006. Article (CrossRef Link)
[33] A. Argyriou, C. A. Micchelli, and M. Pontil, "Learning convex combinations of continuously parameterized basic kernels," in *Proc. of 18th Annu. Conf. Learn*. Theory, Bertinoro, Italy, pp. 338–352, 2005. Article (CrossRef Link)
[34] C. S. Ong, A. J. Smola, and R. C. Williamson, "Learning the kernel with hyperkernels," *J. Mach. Learn. Res.*, vol. 6, pp. 1043–1071, 2005. Article (CrossRef Link)
[35] 25. A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proc. of 24th Int. Conf. Mach.Learn.*, Corvallis, OR, pp. 775–782, 2007.

Article (CrossRef Link)
[36] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008. Article (CrossRef Link)
[37] Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large scale multiple kernel earning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006. Article (CrossRef Link)
[38] A. Zien and C. S. Ong, "Multiclass multiple kernel learning," in *Proc. of 24th Int. Conf. Mach. Learn.*, Corvallis, OR, pp. 1191–1198, 2007. Article (CrossRef Link)
[39] C. J. C. Burges, "Simplified support vector decision rules," in *Proc. of 13th Int. Conf. Mach. Learn.*, San Mateo, CA, pp. 71–77, 1996. Article (CrossRef Link)
[40] B. Scholkopf and A. Smola, "Learning with Kernels," *Cambridge,MA: MIT Press*, 2002. Article (CrossRef Link)
[41] D. Nguyen and T. Ho, "An efficient method for simplifying support vector machines," in *Proc. of 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, pp. 617–624, 2005. Article (CrossRef Link)
[42] M. Wu, B. Scholkopf, and B. Bakir, "A direct method for building sparse kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 7, pp.603–624, 2006. Article (CrossRef Link)
[43] M. Wu, B. Scholkopf, and G. Bakir, "Building sparse large margin classifiers," in *Proc. of 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, pp. 996–1003, 2005. Article (CrossRef Link)
[44] Jun Yin, Zhong Jin, "Kernel sparse representation based classification," *Neurocomputing*, vol. vol. 77, no. 1, pp. 120-128, 2012. Article (CrossRef Link)
[45] Shenghua Gao, Ivor Wai-Hung Tsang, Liang-Tien chia, "Kernle Sparse Representation for Image Classification and Face Recognition," In *Proc. of Proceedings 11th European Conference on Computer Vision*, pp. 1-14, 2010. Article (CrossRef Link)
[46] Li Zhang, Wei-Da zhou, "Kernel sparse representation-based classifier," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1684-1695, 2012. Article (CrossRef Link)
[47] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996. Article (CrossRef Link)
[48] Lanckriet, G.R.G., et al., "Learning the Kernel Matrix with Semidefinite Programming," *J.Machine Learning Research 5*, 27–72, 2004. Article (CrossRef Link)
[49] P. J. Phillips, H. Moon, S. A. Rivzi, and P. Rauss, "The FERET EvaluationMethodology for Face-Recognition Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, 2000. Article (CrossRef Link)

Hao Zheng received the his BS degree from SouthEast University in 1998, the MS degree from Nanjing University Posts and Telecommunications in 2005, and the PhD degree in pattern recognition and intelligence system from Nanjing University of Science and Technology in 2013. He visited the Center of Quantum Computation & Intelligent Systems, University of Technology Sydney, Australia, from September 2013 to March 2014. He is currently an associate professor with the School of Mathematics and Information Technology at the Nanjing Xiaozhuang University. His research interests include pattern recognition, image processing,face recognition,computer vision.

Qiaolin Ye received the BS degree in Computer Science from Nanjing Institute of Technology, Nanjing, China, in 2007, the MS degree in Computer Science and Technology from Nanjing Forestry University, Jiangsu, China, in 2009, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology, Jiangsu, China, in 2013. He is currently an associate professor with the computer science department at the Nanjing Forestry University, Nanjing, China. He has authored more than 30 scientific papers in pattern recognition, machine learning and data mining. His research interests include machine learning, data mining, and pattern recognition.

Zhong Jin received his BS in mathematics, MS in applied mathematics,and PhD in pattern recognition and intelligence systems from Nanjing University of Science and Technology (NUST), China,in 1982, 1984, and 1999, respectively. He is a professor in the Department of Computer Science, NUST, and previously was a research assistant at the Department of Computer Science and Engineering, Chinese University of Hong Kong from 2000 to 2001.He visited the Laboratoire HEUDIASYC, Universite de Technologiede Compiegne, France, from October 2001 to July 2002. He visited the Centre de Visio per Computador, Universitat Autonoma de Barcelona, Spain, as the Ramon y Cajal Research Fellow from September 2005 to October 2005. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis, and content-based image retrieval.