Arab J Sci Eng (2018) 43:7113–7125
https://doi.org/10.1007/s13369-017-2952-x

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Weighted Feature Space Representation with Kernel for Image Classification

Yongbin Qin[1,2] · Chunwei Tian[3]

Received: 13 October 2016 / Accepted: 10 October 2017 / Published online: 27 November 2017

Abstract The kernel method is a very effective and popular method to extract features from data such as images. A novel method is presented to enhance traditional kernel method for face image representation in this paper, which is very suitable to treat the high-dimensional datasets. The proposed method is called weighted kernel representation-based method (WKRBM) in this paper. WKRBM assumes that the test sample can be expressed by all the training samples and linear solution in the mapping space. It uses the obtained linear combination to recognize face images. In particular, the coefficients of a linear combination can be set as the optimal weight that is an important factor to obtain better performance for image classification. The rationale, characteristics, and advantages of the proposed method are presented. The analysis describes that WKRBM outperforms collaborative representation-based kernel method for image recognition. Extensive experimental results illustrate that WKRBM has partial properties of sparsity, which is effective to recognize images.

Keywords Image classification · Representation-based kernel · Adaptive weights with kernel · Feature space representation

✉ Yongbin Qin
ybqin@foxmail.com

[1] School of Computer Science and Technology, Guizhou University, Guiyang 550025, China

[2] Guizhou Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

[3] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China

1 Introduction

The kernel method has been widely applied for feature extraction and classification in pattern recognition and image processing [1–3]. Kernel regression, kernel Fisher discriminant analysis, and kernel principal component analysis (KPCA) are typical kernel methods and have a lot of applications. If we set an input space, a kernel method would finish its work in a novel space (i.e., feature space or mapping space) in general, which stems from original input space. One of the characteristics of the kernel method is that it does not use the virtue of kernel functions to implement details of the corresponding transform. As a consequence, the kernel method would produce lower computational cost than conventional nonlinear methods. One of the merits of kernel methods is that it cannot generate error classification when it is used to represent the test sample. The reason of problems above can be showed briefly as follows. The sparse method is that only uses training samples and sparse solution to express a test sample. In other words, this method presumes that the elements of spare solution are 0 or near 0 when it can exploit training samples and sparse solution to express the test sample. The total of the training samples is smaller than the dimension of samples in general in face image classification and the test sample is represented by Refs. [4,5] must emerge much representation error. Representation error usually increases the probability of false classification result of the test sample. And conventional sparse method also produces expensive computational cost. That is, the kernel method is a good choice for image representation.

In practice some scholars have proposed different methods to address the problems of incorrect representation for face recognition and accelerate the classification process of kernel-based methods [6–9]. Billings et al. [10] proposed to fuse the orthogonal least-squares algorithm and kernel nonlinear discriminant analysis in improvement classification speed. The proposed algorithm only uses a small portion of training samples to express and sort the test sample, which improves the classification efficiency. However, this method of Ref. [11] does not have any regularization term, which may make orthogonal least-squares algorithm over-fitting. Cawley et al. [11] proposed another kernel method to deal with the problem of image classification, which used the kernel trick to get a nonlinear variant of Fisher's analysis methods. This method exploits leave-one-out cross to improve kernel Fisher discriminant analysis, which can speed up the processing and decrease the rate of classification errors on image classification. However, kernel Fisher discriminant might incur the ill-posed phenomenon in its real applications [12,13]. Some scholars proposed a lot of regularization methods to address this problem. Yang [14] proposed to integrate PCA plus LDA method and Fisherface [15] to tackle difficult question of kernel Fisher discriminant above. Baudat et al. [16] used the QR decomposition algorithm to eliminate the zero eigenvalues and keep away from singularity. Unfortunately, these methods dropped important discriminant information, such as the within-class covariance matrix in the null space which was very useful to address the difficult problem of small sample size [17–21]. Lu et al. [22] utilized generalization of direct-LDA [19] to solve the size problem of appeared small sample, which obtained good performance for face recognition. However, when different test samples were expressed by linear coefficients and training samples for this method, it had different kernel methods with the same linear combination coefficients, which cannot properly represent test samples [22]. As consequence, the improvement of kernel method is an effective method to solve problems above [23].

This paper fully considered the diversity of different samples and proposed a novel method WRKBM (weighted kernel representation-based method) to deal with the drawbacks of general kernel method, which can use different kernel methods and varying linear combination coefficients rather than different kernel method and the same linear combination coefficients to properly represent different test samples. Meanwhile, our method has partial sparsity, which is important and effective for image recognition [24]. The main idea of WRKBM is that employs obtained all the training samples to express obtained the test sample and construct the residual to classify the test sample from new mapping space is proposed in the paper. The implementations of WKRBM have the following steps. Firstly, we use the Gaussian function method to obtain the kernel representations in the old mapping space for images, which is corresponding to test samples and training samples from original data. Secondly, the weights of coefficients of the linear representation can be obtained by WKRBM in the new mapping space. Finally, we utilize the obtained weights and kernel function to establish the relation of the test sample with all training samples. Our proposed method is regarded as weighted kernel representation-based method (WKRBM). We also analyze the rationale, characteristics and advantages of WKRBM and the differences between WKRBM and conventional kernel methods to show its performance. The experimental results illustrate that WKRBM not only can have high accuracy for image classification, but also possess partial attribute of sparsity; that is, the sparse solution is near 0 or equal 0. Sparseness is effective for image classification [6]. To make the proposed method obtain better performance, WKRBM is used to compute the weights which make different classes properly represent a test sample. A further merit of WKRBM is that it can decrease the dimension of the original image. For example, an original representation-based method should use a m dimensional vector to represent a sample. If m is very large, then the computational complexity will be very high. However, WKRBM can be easily implemented and has low computational cost.

The structure of this paper is as follows. Section 2 illustrates WKRBM method; Sect. 3 presents the rationale, characteristics, advantages of WKRBM and differences between WKRBM and KRBM, differences between WKRBM and sparse method. Section 4 describes experimental results of massive experiments. Section 5 provides the conclusion of this paper.

2 WKRBM Method

WKRBM includes two main steps. The first step uses weighted linear representation of all the training samples to express the test samples in the new mapping space. The second step employs the obtained representation results to sort the test sample. The specific implementations of WKRBM are presented as follows. Suppose A denotes all the training samples from original data, where $A \in [A_1, A_2, \ldots, A_n]$ and n denotes total of all the subjects. We assume that a given test sample is represented by Y in the original space. We employ the nonlinear mapping ϕ to obtain a new space (i.e., old feature space or old mapping space) from the original sample. In the old mapping space, the ith training sample A_i from the original space can be denoted by $\phi(A_i)$. We assume that the given test sample $\phi(Y)$ can be linearly represented by training samples $\phi(A_i)$ and coefficients in the obtained old mapping space, where $i \in [1, 2, 3, \ldots, n]$ and the formula $\phi(Y) = \underbrace{\beta_1\phi(A_1) + \beta_2\phi(A_2) \cdots + \beta_n\phi(A_n)}_{n}$ is intuitively used to denote the relation of them. We assume that all the samples are column vectors. And the formula above is transformed into formula (1).

$$\phi(Y) = \xi\delta \tag{1}$$

where $\xi = [\phi(A_1), \phi(A_2), \ldots, \phi(A_n)]$ and $\delta = (\beta_1, \beta_2, \ldots, \beta_n)^{\mathrm{T}}$ is the coefficients of the linear representation of all the training samples ξ in the old mapping space. Because ϕ is unknown and ξ is not a square matrix, we cannot directly obtain δ. However, $\xi^{\mathrm{T}}\xi$ is a square matrix, and we can convert formula (1) into formula (2) to solve δ.

$$\xi^{\mathrm{T}}\phi(Y) = \xi^{\mathrm{T}}\xi\delta \tag{2}$$

We assume that kernel function $k(A_i, A_j) = \phi^{\mathrm{T}}(A_i)\phi(A_j)$ [9,25,26]. We can further convert Eq. (2) into Eq. (3).

$$K_Y = K\delta \tag{3}$$

In particular, K_Y and K, respectively, represent kernel function of the test sample and kernel function of the training sample in the mapping space (i.e., old mapping space). K_Y,K, and δ are defined as follows:

$$K_Y = [k(A_1, Y), k(A_2, Y), \ldots, k(A_n, Y)]^{\mathrm{T}},$$
$$K = [k_1, k_2, \ldots, k_n] \text{ and}$$
$$k_i = [k(A_1, A_i), k(A_2, A_i), \ldots k(A_n, A_i)]^{\mathrm{T}}$$

where $i \in [1, 2, 3, \ldots, n]$.

We can obtain the solution of Eq. (3) by Eq. (4), where μ is a nonnegative and nonzero constant and E is the identity matrix.

$$\delta = \begin{cases} K^{-1}K_Y, & K \text{ is not singular} \\ (K + \mu E)^{-1}K_Y, & K \text{ is singular} \end{cases} \tag{4}$$

It is obvious that $K_Y = (k_1, k_2, \ldots, k_n)\delta = \beta_1 k_1 + \cdots + \beta_n k_n$, where $k_j = [k(A_1, A_j), k(A_2, A_j), \ldots, k(A_n, A_j)]^{\mathrm{T}}$ and $j \in [1, 2, \ldots, n]$. This illustrates that a test sample k_y is expressed in term of k_i in the old space. And, k_i is the ith training sample from the old mapping space. Meanwhile, previous study makes us know that the contribution of each training sample for expression of test sample in the mapping space is important for classification. Thus, the issue of improving the accuracy on face recognition might be transformed into a new issue of reasonably setting the weights.

In response to phenomena above, we proposed the following scheme with the purpose of adaptively obtaining the coefficients of the linear combination in the mapping space method via kernel functions. We obtain a new linear combination in the new mapping space, which is the basis of the old mapping space. We assume that ϕ' is new nonlinear mapping in the new mapping space. We assume that the obtained test sample $\phi'(Y)$ is expressed by all the input image training samples $\phi' = [\phi'(A_1), \phi'(A_2), \ldots, \phi'(A_n)]$ in the new mapping space, i.e., $\phi'(Y) = \sum_{i=1}^{n} \beta'_i \phi'(A_i)$ can be obtained. Let $\delta' = [\beta'_1, \beta'_2, \beta'_3, \ldots, \beta'_n]^{\mathrm{T}}$. Finally, the WKRBM method obtains the solution of $\phi'(Y) = \sum_{i=1}^{n} \beta'_i \phi'(A_i)$.

In order to implement operations above, we first obtain the deviation d_i of the ith training sample and test sample in the obtained old mapping space and use it as weight of $\phi(A_i)$ by Eq. (5).

$$d_i = \|K_Y - k_i\delta_i\|_2 \tag{5}$$

where $k(A_i, A_j) = (\phi(A_i))^{\mathrm{T}}\phi(A_j)$. Especially, K'_Y and K' represent kernel function of the test sample and kernel function of the training sample in the new mapping space. The meanings of K'_Y and K' are as follows:

$$K_Y = [k(A_1, Y), k(A_2, Y), \ldots, k(A_n, Y)]^{\mathrm{T}},$$
$$K = [k_1, k_1, \ldots, k_n] \text{ and}$$
$$k_i = [k(A_1, A_i), k(A_2, A_i), \ldots k(A_n, A_i)]^{\mathrm{T}}$$

where $i \in [1, 2, 3, \ldots, n]$.

Second, we use the product of the obtained weight of Eq. (5) and old mapping $\phi(A_i)$ to obtain the new mapping $\phi'(A_i)$ of the ith training sample in the new mapping space. The new mapping $\phi'(A_i)$ of the ith training sample in the new mapping space will be represented as Eq. (6):

$$\phi'(A_i) = d_i\phi(A_i) \tag{6}$$

Third, according to illustration of the kernel function and Eq. (6) above, the new kernel function in the new mapping space can be obtained.

$$\begin{aligned} k'(A_i, A_j) &= (\phi'(A_i))^{\mathrm{T}}\phi'(A_j) = (d_i\phi(A_i))^{\mathrm{T}}(d_j\phi(A_j)) \\ &= d_i d_j k(A_i, A_j) \end{aligned} \tag{7}$$

Then, an equation $k'_Y = k'\delta'$ is solved by Eq. (8).

$$\delta' = \begin{cases} (k')^{-1}k'_Y, & k' \text{ is not singular} \\ (k' + \mu E)^{-1}k'_Y, & k' \text{ is singular} \end{cases} \tag{8}$$

where k' and $k'_Y = (d_i k) \times k_Y$, respectively, denote the matric and vector corresponding to the new kernel functions of training samples and test sample in the new mapping space.

Finally, we can use the test sample from the old mapping space, obtained coefficients and training samples of one subject from the new mapping space to construct the residual of every class and exploit the residual to sort test sample. We assume that all the training samples of the pth subject are $k'' = [k'_1, k'_2, \ldots, k'_l]$ and l denotes the number of training samples of each subject. Meanwhile, we assume that $\delta'' = [\beta'_1, \beta'_2, \ldots, \beta'_n]^{\mathrm{T}}$ is the vector associated with class p. Let residual of the pth class be e_p and e_p is calculated by Eq. (9).

$$e_p = \left\|k'_Y - k''\delta''\right\|_2^2 \tag{9}$$

When a subject has the minimum error, WKRBM thinks that test sample k_Y belongs to this subject. In other words, if the test sample is close to the linear representation of all the training samples of the pth subject, WKRBM thinks that the test sample belongs to this subject.

Classification is determined below if $t = \underset{p}{\operatorname{argmin}}\, e_p$, the test sample is considered to be the pth class.

$$t = \underset{p}{\operatorname{argmin}}\, e_p \tag{10}$$

3 Analysis of the Proposed Method

We will introduce the rationale of WKRBM, illustrate its characteristics, show its advantages, and provide the differences from WKRBM and other methods in this section.

3.1 Rationale, Characteristics, and Advantages of WRKBM

WKRBM has three characteristics and merits. First, it can obtain a linear system based kernel method. Second, it can adaptively obtain weights by using the obtained kernel function and the weights to allow test samples to be better recognized. The weight is a proper penalized factor on training samples. It allows effects on the test sample of different training samples to be better exploited. The idea to employ the difference from training sample and the test sample as a weight is partially similar with the scheme in [30], but the implementation procedure is very different. Third, WKRBM has higher accuracy for image recognition than conventional kernel method. The high accuracy is shown in Sect. 4. WKRBM can also decrease the dimension of original images in the mapping space, which increases the flexibility of the operation and improves the effectiveness of image processing. WKRBM also has partial properties of sparsity. The sparsity is effective for image classification [30]. The sparsity of WKRBM is better than KRBM shown as in Figs. 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25 and 26. WKRBM can be easily implemented.

The rationale of WKRBM is described in detail as follows. We assume that above nonlinear mapping ϕ is known, ϕ can be solved by Eq. (2). It is obvious to obtain that $\xi^{\mathrm{T}}\phi(Y) = [\phi^{\mathrm{T}}(A_1)\phi(Y), \phi^{\mathrm{T}}(A_2)\phi(Y), \ldots, \phi^{\mathrm{T}}(A_n)\phi(Y)]^{\mathrm{T}}$, where $\phi(A_i)$ and $\phi(Y)$, respectively, denote the ith training sample and test sample from the old mapping space and $(\xi^{\mathrm{T}}\xi)_{ij} = \phi^{\mathrm{T}}(A_i)\phi(A_j)$,$i.j \in [1, 2, 3, \ldots, n]$. According to the vector knowledge, we know that $\phi^{\mathrm{T}}(A_i)\phi(Y) = \|\phi(A_i)\| \bullet \|\phi(Y)\| \cos\theta_i$, where θ_i means the angle between vectors $\phi(A_i)$ and $\phi(Y)$. When all training samples are unit vectors in the old mapping space, $\xi^{\mathrm{T}}\phi(Y) = [\phi^{\mathrm{T}}(A_1)\phi(Y), \phi^{\mathrm{T}}(A_2)\phi(Y), \ldots, \phi^{\mathrm{T}}(A_n)\phi(Y)]^{\mathrm{T}}$ can convert an issue of the cosine similarity between each training sample and the test sample from old mapping space to solve the problem Eq. (2). If two training samples are orthogonal in the obtained old mapping space, $\xi^{\mathrm{T}}\xi$ would be the unit matrix. We can obtain the solution of Eq. (2), which is $\delta = \xi^{\mathrm{T}}\phi(Y) = [\phi^{\mathrm{T}}(A_1)\phi(Y), \phi^{\mathrm{T}}(A_2)\phi(Y), \ldots, \phi^{\mathrm{T}}(A_2)\phi(Y)]^{\mathrm{T}} = [\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_n]^{\mathrm{T}}$. Similarly, we can obtain the coefficient solution δ' of the new mapping space $\delta' = [(\phi'(A_1))^{\mathrm{T}}\phi'(Y), (\phi'(A_2))^{\mathrm{T}}\phi'(Y), \ldots, (\phi'(A_n))^{\mathrm{T}}\phi\prime(Y)]^{\mathrm{T}} = \left[\cos\theta'_1, \cos\theta'_2, \ldots, \cos\theta'_n\right]^{\mathrm{T}}$, where $\phi'(A_i)$ and $\phi'(Y)$, respectively, denote the ith training sample and test sample from new mapping space and θ'_i means the angle between vectors $\phi'(A_i)$ and $\phi'(Y)$. This shows that the solution vector $\delta\prime = [\beta'_1, \beta'_2, \beta'_3, \ldots, \beta'_n]^{\mathrm{T}}$ of WKRBM makes up with n components that are used to express the difference of the test sample and training samples in the obtained new mapping space. As shown in Ref. [25], if the difference between the ith training sample and the test sample is smaller, β' is usually large. In other words, the contribution of the ith training sample (i.e., $\beta'_i k'_i$) is large. If difference between a sample from the training samples of the ith subject and the test sample is little, test sample is considered to be this subject. We use Eq. (9) to find the subject, which is the most similar to this test sample and the test sample is classified into this subject by Eq. (10). This partially illustrates the rationale of WKRBM. As a consequence, WKRBM is proper.

3.2 The Differences of WKRBM and Other Methods

In Sect. 3.2, first, conventional sparse method [5] with WKRBM on the sample representation will be compared. Then, we will show the differences between conventional kernel method and WKRBM and between conventional sparse method and WKRBM. We know that Ref. [5] only used training samples and sparse solution to represent the test sample. In other words, the above sparse representation method attempts to make most elements of sparse solution are near or equal to 0 [27]. All training samples can be utilized to express the test sample and classify it in a certain condition, which does not cause any error in theory [28]. Unfortunately, this condition usually cannot be reached and this representation method would bring the classification error. If total of training samples is small, the error of classification would get greater. Obviously, face recognition problem is not a low-dimensional problem, where the total of training samples is smaller than the dimensionality of samples. In the processing of image processing, among training samples usually have certain relations, which might increase the probability of error classification of this test sample. WKRBM can effectively use the nonlinear mapping to deal with the correlated problem in the mapping space and extract more important features. WKRBM may better express the test sample than

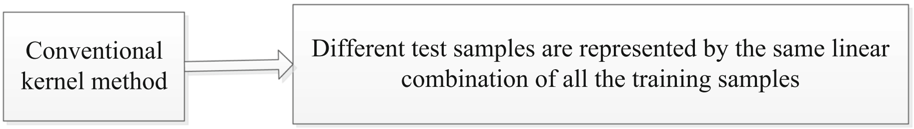
Conventional kernel method
Different test samples are represented by the same linear combination of all the training samples

Fig. 1 Characteristic of the conventional kernel method

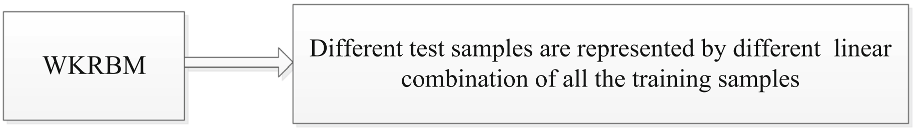
WKRBM
Different test samples are represented by different linear combination of all the training samples

Fig. 2 Characteristic of WKRBM

conventional sparse representation methods. It is notable that kernel method represents test sample and classify it in the mapping space has been proposed [29].

WKRBM is a kernel-based representation method. To better show the performance of WKRBM, we analyze the differences of WKRBM and conventional kernel methods. Here kernel-based representation methods mainly refer to kernel principle component analysis (KPCA) [30] and kernel Fisher discriminant analysis (KFDA) [31]. Above methods are typical kernel methods, which can use kernel functions to map original test samples and original training samples from original data into obtained mapping space. Thus, each test sample will be expressed by n kernel functions, where total of all the training samples is n. This kind of kernel methods indeed uses varying kernel functions corresponding to the same linear combination to represent varying test samples. However, these test samples use the same coefficients feature extractor in the mapping space.

The representation of the test sample is different in WKRBM method. First, the test sample is mapped into another space (old feature space or old mapping space). Then, it uses kernel function and the property of representation in this mapping space to compute weights of coefficients of linear representation. Finally, the test sample is expressed and classified by coefficients and training samples in new obtained mapping space. The implementations of WKRBM are as follows in detail.

Step1 Divide all the original images into two sets, i.e., the set A of training samples and set of Y' test samples. In particular, it denotes a given test sample Y.

Step 2 Employ the nonlinear mapping ϕ to obtain the ith training sample $\phi(A_i)$ and the test sample $\phi(Y')$ in the old mapping space. Utilize equation $k(A_i, A_j) = \phi^{\mathrm{T}}(A_i)\phi(A_j)$ and Eq. (3) to compute kernel functions K_Y and K corresponding to the test sample and training sample in the old mapping space, respectively.

Step 3 Obtain the linear combination coefficients δ of equation $K_Y = K\delta$ by Eq. (4) in the old feature space.

Step 4 Obtain deviation d_i of the ith sample by Eq. (5).

Step 5 Exploit the product of d_i and the ith training sample to obtain new mapping relation δ' of the new feature space by Eq. (6).

Step 6 Calculate the training samples and test samples by equation $k'_Y = (d_i k) \times k_Y$ and Eq. (7) in the new feature space, respectively.

Step 7 Use Eq. (8) to solve the coefficients of the linear combination in the new feature space.

Step 8 Obtain the residual of the pth class by Eq. (9).

Step 9 Classify the test sample by Eq. (10).

Previous study makes us know that, we can use Eqs. (4), (5), (6), (7), and (8) to compute the coefficients of the linear combination in new obtained mapping space. When the linear system is established, we also use $g_p = \beta'_i k'_i + \beta'_{i+1} k'_{i+1} + \cdots + \beta'_{i+l} k'_{i+l}$ to classify test sample Y. It is obvious that, β'_i and $k'_i (i \in [1, 2, 3, \ldots, n])$ vary with test sample Y. That is, kernel function and obtained coefficients vary with the test sample in the new mapping space. This is obvious difference with kernel methods. Meanwhile, we design a novel algorithm to obtain special linear combination in the mapping space, which makes it have ability to better express a given test sample and classify it. To intuitively show the difference of conventional kernel method and WKRBM, we provide Figs.1 and 2 below.

4 Experiments and Results

In Sect. 4, this paper designs massive experiments to inspect the capability of WKRBM on ORL [32], AR [33], and GT [34] databases. As shown later, the feasibility and good performance of WKRBM will be described by these experiments. In the experiments, two-dimensional principle component analysis (2D-PCA), naïve collaborative representation classification (CRC), kernel representation-based method (KRBM) (i.e., Ref. [25] proposed this method) [25], fast iterative shrinkage thresholding algorithm (FISTA) [35] and L1-regularized least-squares (L1LS) [36] are used as comparative experiments to show the performance of WKRBM. 2DPCA is one of typical traditional methods to extract features [37]. 2DPCA directly uses two-dimensional original image to construct the covariance matrix and employs its eigenvectors to extract image feature. PCA (principal component analysis) [38] need to convert two-dimensional original image into one-dimensional vector, which results in PCA has higher computational cost than 2DPCA to extract feature for image processing. That is, PCA not only has higher computational cost, but also ignores the relation between different vectors, which loses some important information. As a consequence, 2DPCA is reasonable for its excellent performance on image recognition. In our experiments, the number of feature vector of 2DPCA is 22 in the experiment. CRC is a symbol of conventional sparse method with 2-norm sparse solution. The main idea of CRC is as follows. First, CRC exploits all the training samples to represent the test sample and the coefficients of the linear combination can be obtained. Then, CRC uses all the training samples of every class and test sample to construct the residual. Finally, CRC utilizes the minimum residual to find the class of the test sample. CRC has tough performance and it has lower computational cost than L1LS and FISTA. In our experiments, let parameter λ be 0.01 when the solution of the coefficients of CRC can be solved by Eq. (9) in Ref. [5]. L1LS and FISTA are typical methods of sparse method with 1-norm sparse solution. L1LS can exploit the truncated Newton interior point method to address l_1-regularized least-squares problem, which has good performance for image processing. FISTA can optimize the wavelet subbandwidth parameter and deal with the problem of reconstruction. Meanwhile, it can find the rate of multi-step approach convergence. KRBM is a novel kernel method. The main implementations of KRBM method are as follows. First, KRBM constructs a linear system to represent the test sample by all the training samples in new space. Then, KRBM uses the linear representation result to classify the test sample. KRBM has lower computational cost than general kernel method. And it has different linear coefficients for different test sample, which obtain better performance on image recognition. When we design the comparison experiments by KRBM, the Gaussian kernel function $k(x_i, x_j) = \exp\left[\frac{-\|x_i - x_j\|^2}{2\sigma}\right]$ is chosen [25]. And parameter σ is 10^7 in this paper. Thus, these methods are used as the comparative experiments, which is powerful to show good performance of WKRBM. Gaussian kernel function is used to conduct the experiments here. Let σ be 10^7 in the Gaussian kernel function $k(x_i, x_j) = \exp\left[\frac{-\|x_i - x_j\|^2}{2\sigma}\right]$. And μ is 0.01 in Eq. (8) in this paper. Each face database can be separated into training set and test set. We show the performance of our method as follows. We first use WKRBM, 2DPCA, CRC, and KRBM to classify face images on the same face dataset when total of training samples is different. Then, WKRBM is utilized to obtain linear coefficients to show that our method exist partial sparse property. It is clear that sparsity is important and effective to recognize face images. In Tables 2, 3, WKRBM, 2PCA and CRC, CRC and KRBM are shown. Figures 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25 and 26 illustrate that values of obtained coefficients in the new mapping space vary from the number of coefficients on different face datasets. As shown in Tables 1, 2, and 3, WKRBM has good performance on image recognition. Comparative experiments on ORL, AR, and GT face datasets are as follows.

Fig. 3 Five images from ORL dataset

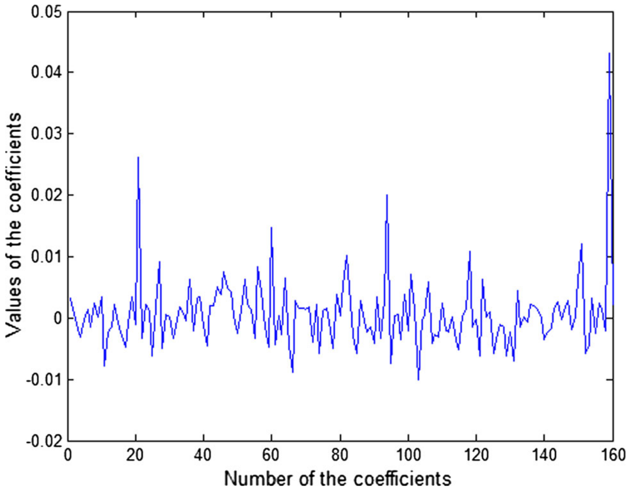

Fig. 4 The obtained coefficients of WKRBM from the ORL face database when the number of training sample of each class is 4

4.1 Experiment on the ORL Dataset

In this section, first 9 images of each class in the ORL dataset are used to test the performance of WKRBM for face recognition. The 360 images are chosen from 40 different persons, and each person takes nine images. Size of each image is 56 × 46 matrix. Different samples of the same subject have different facial expressions, which is an important characteristic. Figure 3 shows 5 different face images from the ORL dataset.

We, respectively, design the experiments when the total of images of each class is 4, 5, and 6. And the first 4, 5 and 6 images of each class are looked upon as training samples and the other images of each class are regarded as test samples. Table 1 describes that error rate of image classification on ORL dataset. This illustrates that WKRBM has higher accuracy than 2D-PCA and CRC, CRC, FISTA, L1LS, and KRBM. For example, WKRBM obtains the rate of classification errors is 52.50, 53.13, 51.67, and 50.00% when the total of training samples of every subject varies from 4 to 7. However, KRBM obtains the rate of classification errors is

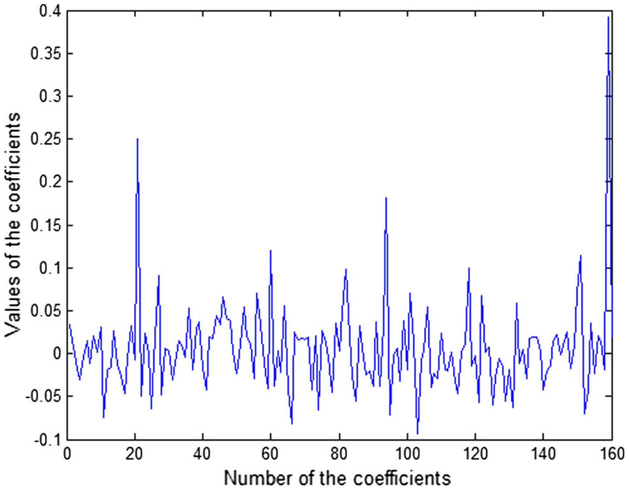

Fig. 5 The obtained coefficients of KRBM from the ORL face database when the number of training sample of each class is 4

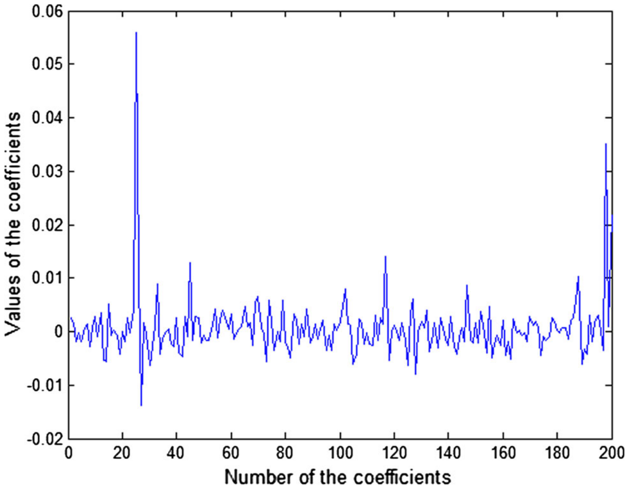

Fig. 6 The obtained coefficients of WKRBM from the ORL face database when the number of training sample of each class is 5

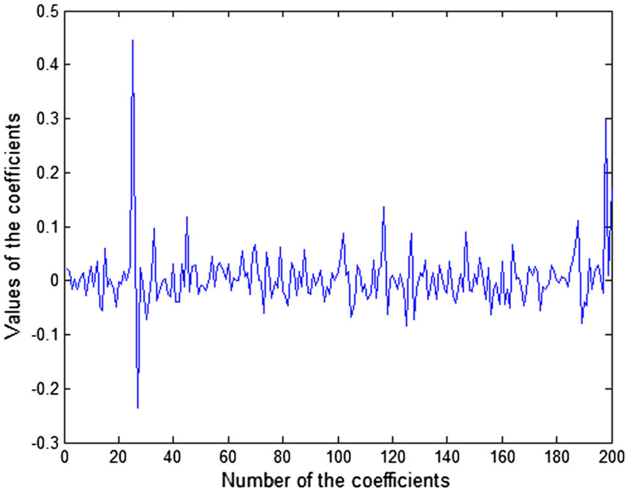

Fig. 7 The obtained coefficients of KRBM from the ORL face database when the number of training sample of each class is 5

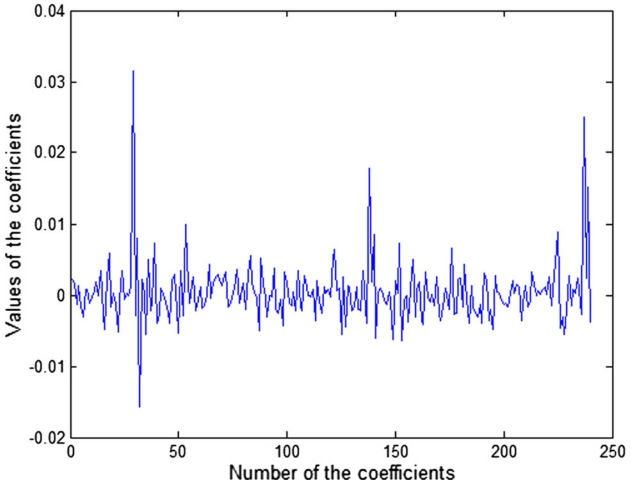

Fig. 8 The obtained coefficients of WKRBM from the ORL face database when the number of training sample of each class is 6

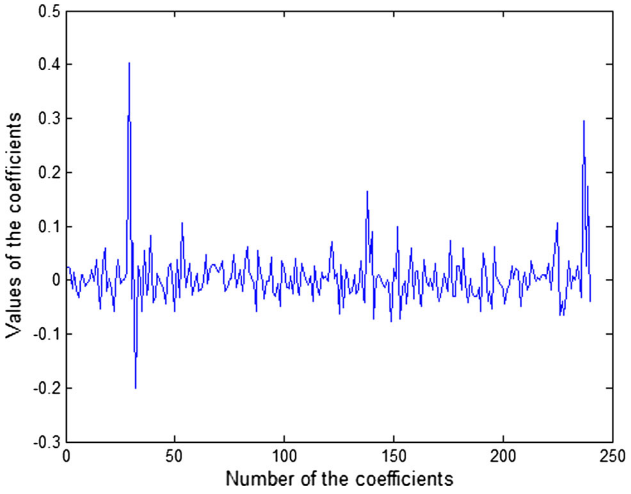

Fig. 9 The obtained coefficients of KRBM from the ORL face database when the number of training sample of each class is 6

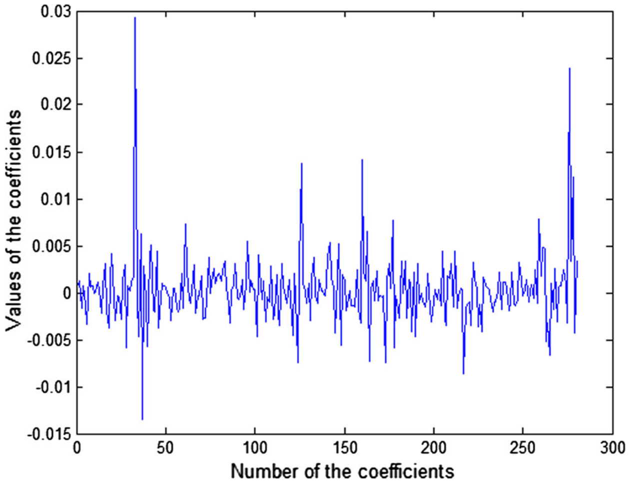

Fig. 10 The obtained coefficients of WKRBM from the ORL face database when the number of training sample of each class is 7

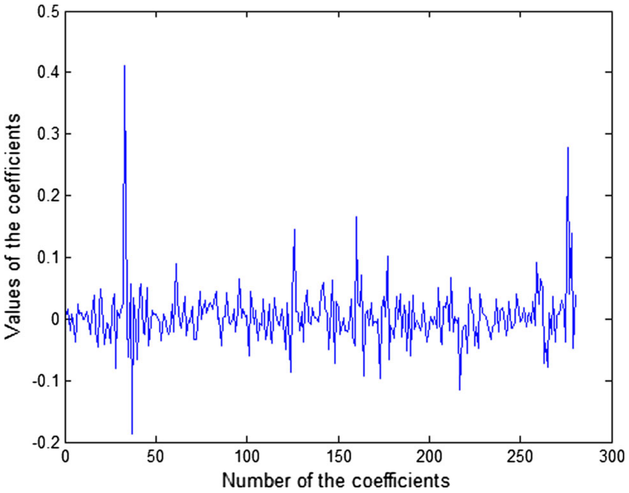

Fig. 11 The obtained coefficients of KRBM from the ORL face database when the number of training sample of each class is 7

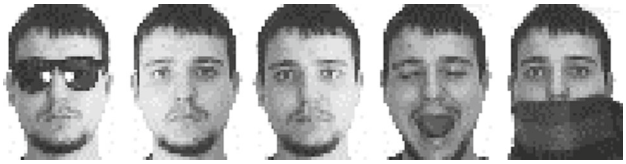

Fig. 12 Five images from AR dataset

53.00, 55.00, 53.33, and 51.25% when the total of training samples of every class is from 4 to 7.

To illustrate the sparsity of coefficients of the linear combination, we choose 4, 5, 6, and 7 images of every subject to be as training samples on ORL face dataset, respectively. As shown in Figs. 4, 5, 6, 7, 8, 9, 10 and 11, we can know that the major coefficients of the linear combination from WKRBM is 0 or nearer 0 than the major coefficients of the linear combination from KRBM when the number of training sample of each class is 4, 5, 6 and 7, respectively. Thus, this shows that WKRBM has partial property of sparsity, which is effective for classification.

4.2 Experiment on the AR Dataset

In this section, first 26 images of each class in the AR dataset are used to test the performance of WKRBM for face recognition. The 3120 images are chosen from 120 persons and each person includes 26 images. Size of each image is 50 × 40 matrix. Different samples of the same subject have different facial expressions, occlusions and illuminations, which are important characteristic. Figure 12 shows 5 different face images from the AR dataset.

We, respectively, design the experiments when the total of images of each class is 14, 15 and 16. And the first 14, 15, and 16 images of each class are looked upon as training samples and the other images of each class are regarded as test samples. Table 2 describes that error rate of image classification on AR dataset. It illustrates that WKRBM has higher accuracy than 2DPCA and CRC, CRC, FISTA, L1LS and KRBM. For example, WKRBM, respectively, obtains error rate of image classification is 7.57, 6.36, and 6.67% when the total of training samples of every subject varies from 14 to 16. However, CRC obtains the rate of classification errors is 14.86, 11.82, and 9.83% when the total of training samples of every class is from 14 to 16.

To illustrate the sparsity of coefficients of the linear combination, we choose 14, 15, and 16 images of every class to be as training samples and only use 50 classes to obtain the coefficients on AR face dataset here, respectively. As shown in Figs. 13, 14, 15, 16, 17, and 18, we can know that the major coefficients of the linear combination from WKRBM is 0 or

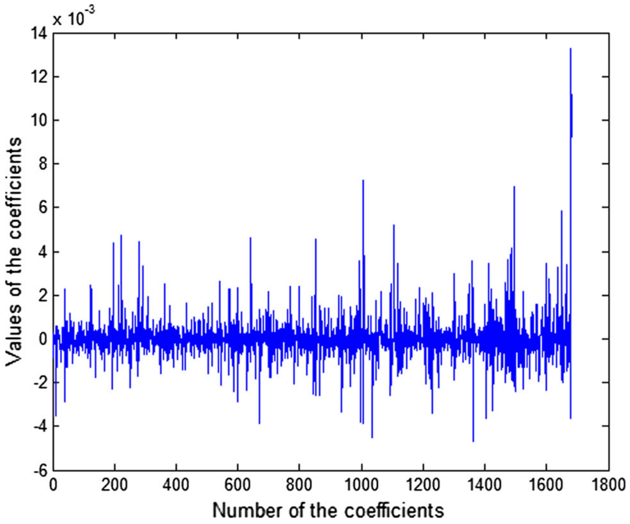

Fig. 13 The obtained coefficients of WKRBM from the AR face database when the number of training sample of each class is 14

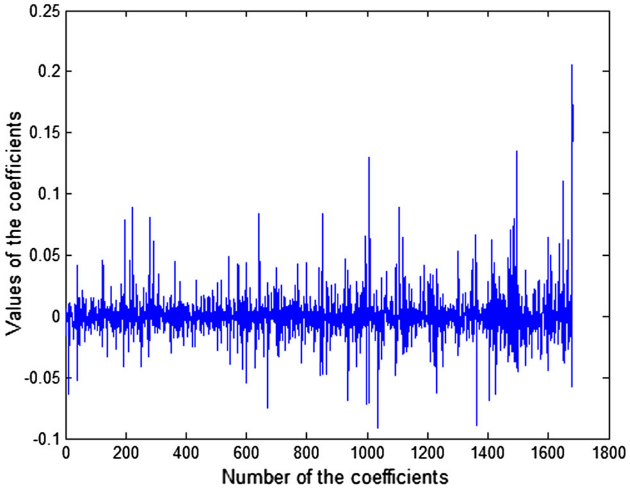

Fig. 14 The obtained coefficients of KRBM from the AR face database when the number of training sample of each class is 14

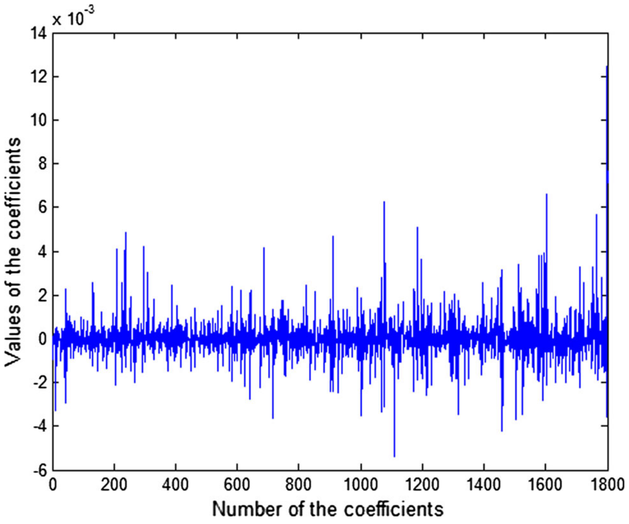

Fig. 15 The obtained coefficients of WKRBM from the AR face database when the number of training sample of each class is 15

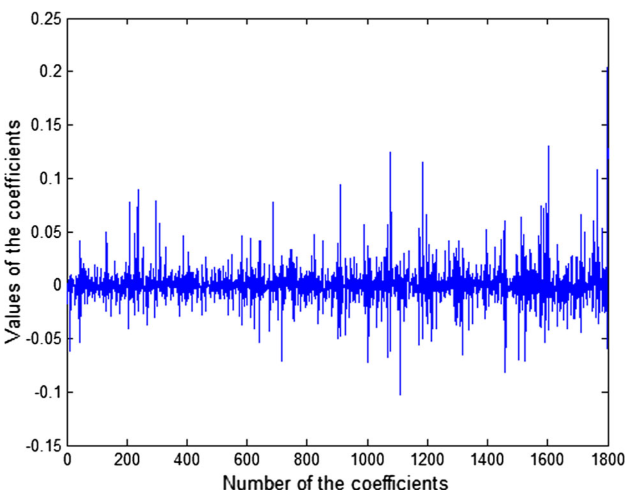

Fig. 16 The obtained coefficients of KRBM from the AR face database when the number of training sample of each class is 15

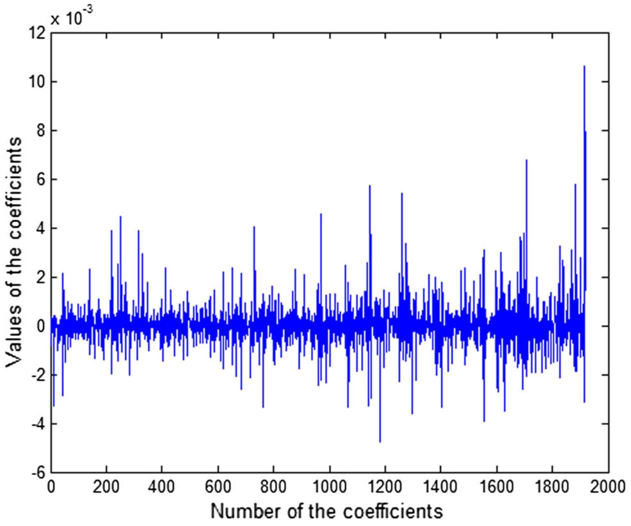

Fig. 17 The obtained coefficients of WKRBM from the AR face database when the number of training sample of each class is 16

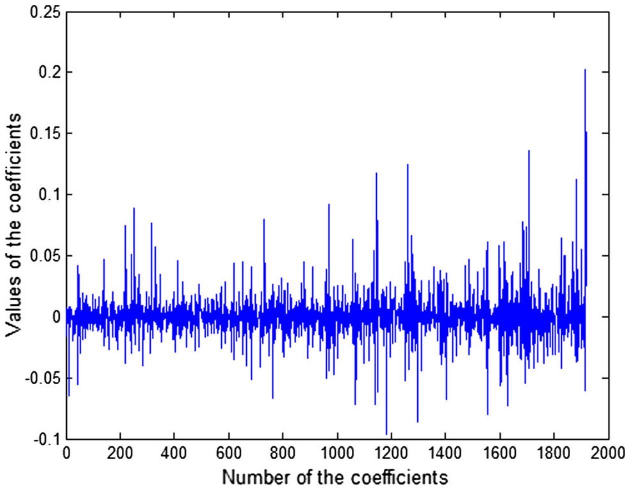

Fig. 18 The obtained coefficients of KRBM from the AR face database when the number of training sample of each class is 16

nearer 0 than the major coefficients of the linear combination from KRBM when the number of training samples of each class is 14, 15, and 16, respectively. Thus, this shows that WKRBM has partial property of sparsity, which is effective for classification.

4.3 Experiment on the GT Dataset

In this section, first 15 images of each class in the GT dataset are used to test the performance of WKRBM for face recognition. The 750 images are chosen from 50 subjects and each subject has 15 images. Size of each image is 40 × 30 matrix. Different samples of the same subject have different facial expressions and illuminations, which are important characteristic. Figure 19 shows 5 different face images from the GT dataset.

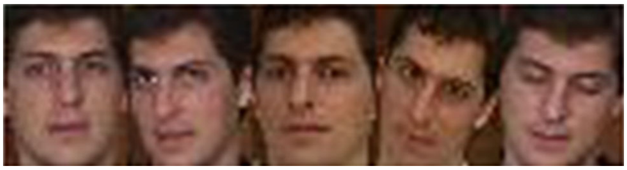

Fig. 19 Five images from GT dataset

We, respectively, design the experiments when the total of images of each class is 3, 4, 5, and 6. And the first 3, 4, 5, and 6 images of each class are looked upon as training samples and the other images of each class are regarded as test samples. Table 3 describes that error rate of image classification on GT dataset. It illustrates that WKRBM has higher accuracy

Table 1 Rate of classification errors (%) on the ORL dataset

Number of training samples every class	4	5	6	7
WKRBM	52.50	53.13	51.67	50.00
Kernel representation-based method (KRBM)	53.00	55.00	53.33	51.25
Naïve collaborative representation classification (CRC)	54.50	56.25	53.33	53.75
Two-dimensional principle component analysis (2DPCA) and CRC	55.00	58.13	61.67	62.50
FISTA	55.00	56.87	55.83	55.00
L1LS	54.00	55.63	51.67	52.50

Table 2 Rate of classification errors (%) on the AR dataset

Number of training samples every class	14	15	16
WKRBM	7.57	6.36	6.67
Kernel representation-based method (KRBM)	8.13	7.20	7.00
Naïve collaborative representation classification (CRC)	14.86	11.82	9.83
Two-dimensional principle component analysis (2DPCA) and CRC	50.00	47.58	44.75
FISTA	21.58	22.20	21.58
L1LS	19.38	16.06	13.25

Table 3 Rate of classification errors (%) on the GT dataset

Number of training samples every class	3	4	5	6
WKRBM	49.50	47.45	44.20	35.56
Kernel representation-based method (KRBM)	51.83	48.00	45.40	37.11
Naïve collaborative representation classification (CRC)	54.67	52.91	51.20	44.40
Two-dimensional principle component analysis (2D-PCA) and CRC	59.67	57.27	54.80	53.11
FISTA	50.38	49.27	48.00	39.11
L1LS	55.33	53.82	51.20	44.44

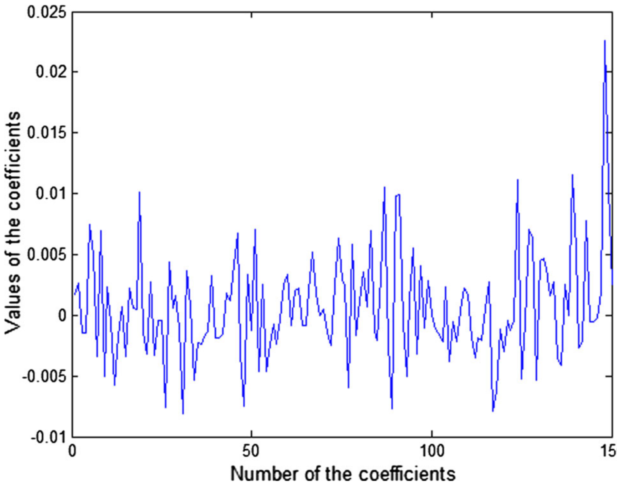

Fig. 20 The obtained coefficients of WKRBM from the GT face database when the number of training sample of each class is 3

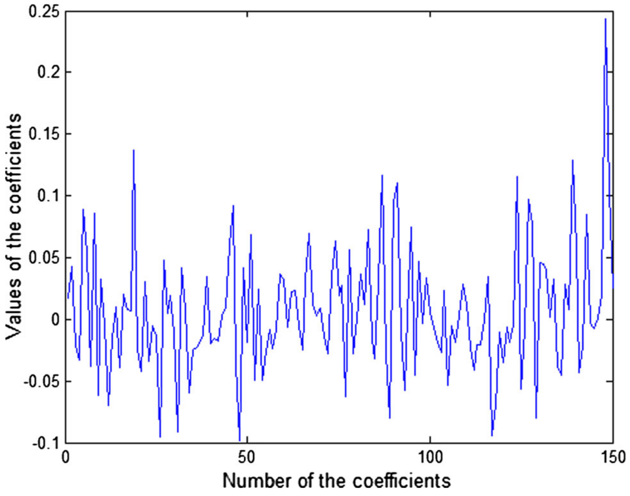

Fig. 21 The obtained coefficients of KRBM from the GT face database when the number of training sample of each class is 3

than 2DPCA and CRC, CRC, FISTA, L1LS and KRBM. For example, the WKRBM obtains the rate of classification errors is 49.50, 47.45, 44.20, and 35.78% when the total of training samples of every subject varies from 3 to 6. However, 2D-PCA and CRC obtains the rate of classification errors is

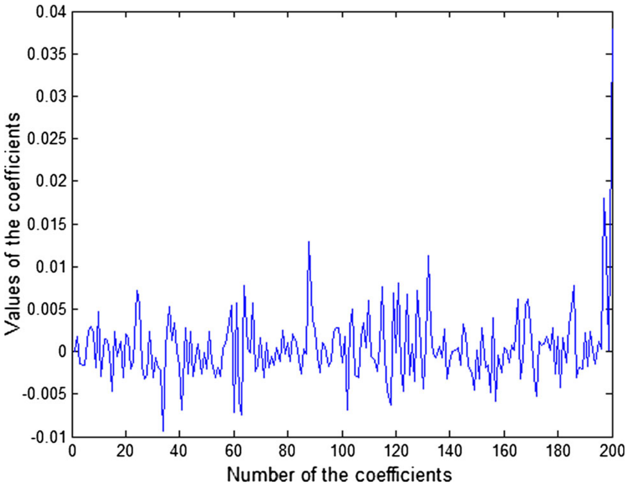

Fig. 22 The obtained coefficients of WKRBM from the GT face database when the number of training sample of each class is 4

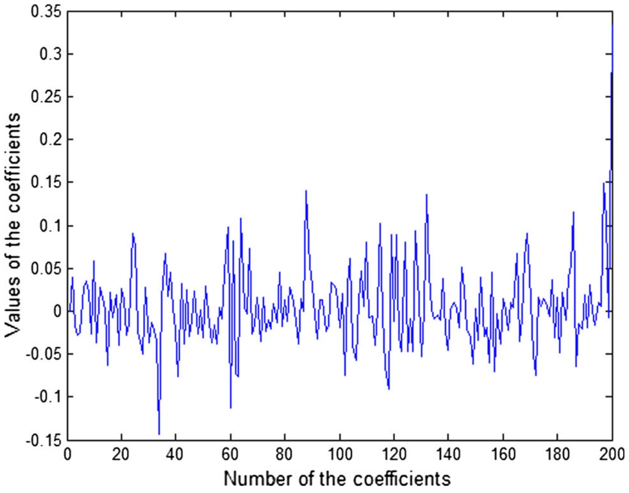

Fig. 23 The obtained coefficients of KRBM from the GT face database when the number of training sample of each class is 4

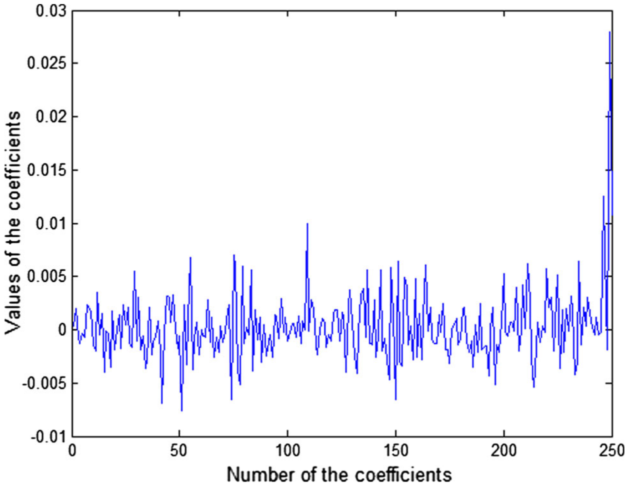

Fig. 24 The obtained coefficients of WKRBM from the GT face database when the number of training sample of each class is 5

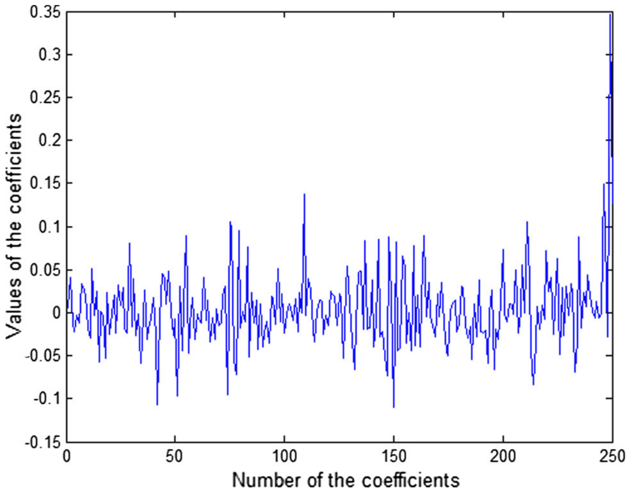

Fig. 25 The obtained coefficients of KRBM from the AR face database when the number of training sample of each class is 5

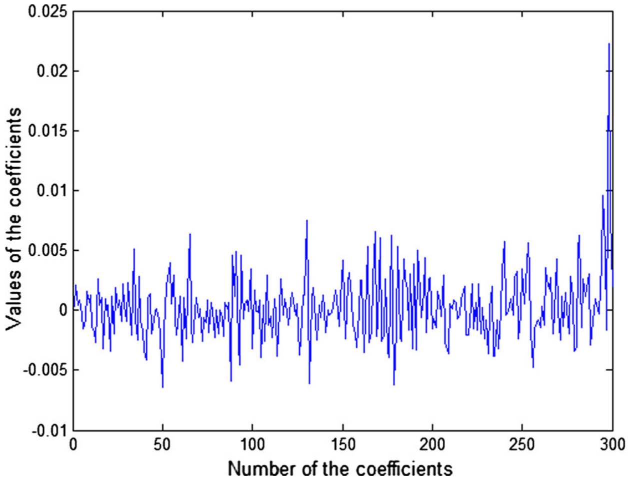

Fig. 26 The obtained coefficients of WKRBM from the GT face database when the number of training sample of each class is 6

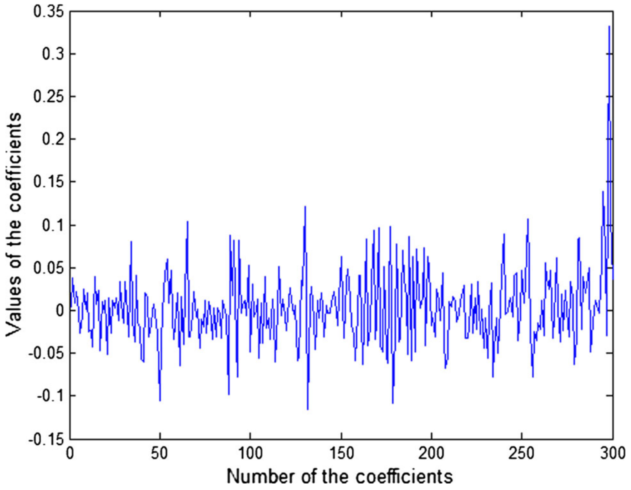

Fig. 27 The obtained coefficients of KRBM from the GT face database when the number of training sample of each class is 6

59.67, 57.27, 54.80, and 53.11% when the total of training samples of every subject is from 3 to 6.

To illustrate the sparsity of coefficients of the linear combination, we choose 3, 4, 5, and 6 images of every class to be as training samples on GT face dataset here. As shown in Figs. 20, 21, 22, 23, 24, 25, 26, and 27, we can know that the major coefficients of the linear combination from WKRBM is 0 or nearer 0 than the major coefficients of the linear combination from KRBM when the number of training sample of each class is 3, 4, 5 and 6, respective. Thus, this shows that WKRBM has partial property of sparsity again, which is effective for image classification.

5 Conclusions

Kernel method and sparse method are effective for image recognition. Traditional kernel method uses linear coefficients and training samples to express different test samples. However, it uses different kernel methods with the same linear combination coefficients, which cannot properly represent test samples. The sparse method is that only uses training samples and sparse solution to express a test sample. It presumes that the elements of spare solution are 0 or near 0 when it can exploit training samples and sparse solution to express the test sample. The total of the training samples is smaller than the dimension of samples in general in face image classification and the test sample is represented that must emerge much representation error. And conventional sparse method also produces expensive computational cost. This paper proposed a novel method WRKBM to overcome defect of the traditional kernel method, which uses different kernel methods and linear combination coefficients to represent different test samples. Meanwhile, this paper shows that WKRBM possesses partial properties of sparsity and it is effective for face recognition.

The main idea of the proposed method is that can be treated as one that uses the weighted sum of all the training samples to represent the test sample by the kernel method and classifies the test sample into the subject, which has the greatest contribution for all the training samples corresponding to the weighted sum.

The analysis shows the rationale, characteristics and advantages. Experimental results show that WKRBM has great performance and has a low error rate for face representation. It is also easy to implement.

Acknowledgements The research work was supported by National Natural Science Foundation of China under Grant No. 61262006 and No. 61540050, Major Applied Basic Research Program of Guizhou Province under Grant No. JZ20142001 and the Science and Technology Foundation of Guizhou Province under Grant No. LH20147636.

References

1. Wang, G.; Shi, N.; Shu, Y.; et al.: Embedded manifold-based kernel Fisher discriminant analysis for face recognition. Neural Process. Lett. **43**(1), 1–16 (2016)
2. Xu, Y.: A new kernel MSE algorithm for constructing efficient classification procedure. Int. J. Innov. Comput. Inf. Control **5**(8), 2439–2447 (2009)
3. Min, H.K.; Hou, Y.; Park, S.; et al.: A computationally efficient scheme for feature extraction with kernel discriminant analysis. Pattern Recognit. **50**, 45–55 (2016)
4. Wright, J.; Ma, Y.; Mairal, J.; et al.: Sparse representation for computer vision and pattern recognition. Proc. IEEE **98**(6), 1031–1044 (2010)
5. Wright, J.; Yang, A.Y.; Ganesh, A.; et al.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009)
6. Scholkopft, B.; Mullert, K.R.: Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX. pp. 41–48 (1999)
7. Xu, Y.; Zhang, D.; Jin, Z.; et al.: A fast kernel-based nonlinear discriminant analysis for multi-class problems. Pattern Recognit. **39**(6), 1026–1033 (2006)
8. Tahir, M.A.; Kittler, J.; Bouridane, A.: Multi-label classification using stacked spectral kernel discriminant analysis. Neurocomputing **171**, 127–137 (2016)
9. Liu, W.; Yu, Z.; Lu, L.; et al.: KCRC-LCD: Discriminative kernel collaborative representation with locality constrained dictionary for visual categorization. Pattern Recognit. **48**(10), 3076–3092 (2015)
10. Billings, S.A.; Lee, K.L.: Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. Neural Netw. **15**(2), 263–270 (2002)
11. Cawley, G.C.; Talbot, N.L.C.: Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognit. **36**(11), 2585–2592 (2003)
12. Weston, J.; Schölkopf, B.; Smola, A.; et al.: Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 623 (2003)
13. Tikhonov, A.N.; Arsenin, V.Y.: Solution of Ill-Posed Problems. Wiley, New York (1997)
14. Yang, M.H.: Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Method. In: Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 215–220 (2002)
15. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
16. Baudat, G.; Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Comput. **12**(10), 2385–2404 (2000)
17. Chen, L.F.; Liao, H.Y.M.; Lin, J.C.; Kao, M.D.; Yu, G.J.: A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit. **33**(10), 1713–1726 (2000)
18. Liu, K.; Cheng, Y.Q.; Yang, J.Y.; et al.: An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method. Int. J. Pattern Recognit. Artif. Intell. **6**(05), 817–829 (1992)
19. Yu, H.; Yang, J.: A direct LDA algorithm for high-dimensional data—with application to face recognition. Pattern Recognit. **34**(10), 2067–2070 (2001)
20. Yang, J.; Yang, J.: Why can LDA be performed in PCA transformed space? Pattern Recognit. **36**(2), 563–566 (2003)

21. Yang, J.; Yang, J.: Optimal FLD algorithm for facial feature extraction. In: Intelligent Systems and Advanced Manufacturing, pp. 438–444. International Society for Optics and Photonics (2001)
22. Lu, J.; Plataniotis, K.N.; Venetsanopoulos, A.N.: Face recognition using kernel direct discriminant analysis algorithms. IEEE Trans. Neural Netw. **14**(1), 117–126 (2003)
23. Fan, Z.; Xu, Y.; Ni, M.; et al.: Individualized learning for improving kernel Fisher discriminant analysis. Pattern Recognit. **58**, 100–109 (2016)
24. Eslami, M.; Jahanshahi, J.A.; Ghorashi, S.A.: Compressive sensing-based PSD map construction in cognitive radio networks. Arab. J. Sci. Eng. **39**(2), 1147–1156 (2014)
25. Xu, Y.; Fan, Z.; Zhu, Q.: Feature space-based human face image representation and recognition. Opt. Eng. **51**(1), 017205-1–017205-7 (2012)
26. Muller, K.R.; Mika, S.; Ratsch, G.; et al.: An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. **12**(2), 181–201 (2001)
27. Xu, Y.; Zhang, D.; Yang, J.; et al.: A two-phase test sample sparse representation method for use with face recognition. IEEE Trans. Circuits Syst. Video Technol. **21**(9), 1255–1262 (2011)
28. Yang, M.; Zhang, L.; Yang, J. et al.: Robust sparse coding for face recognition. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 625–632. IEEE (2011)
29. Chen, S.; Hong, X.; Harris, C.J.: Regression based D-optimality experimental design for sparse kernel density estimation. Neurocomputing **73**(4), 727–739 (2010)
30. Xu, Y.; Zhang, D.; Song, F.; et al.: A method for speeding up feature extraction based on KPCA. Neurocomputing **70**(4), 1056–1061 (2007)
31. Mika, S.; Rätsch, G.; Müller, K. R.: A mathematical programming approach to the kernel fisher algorithm. In: Advances in Neural Information Processing Systems, vol. 13, pp. 591–597 (2001)
32. AT & T Laboratories, "The database of faces," 2002. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (29 December 2011)
33. http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html
34. http://www.face-rec.org/databases/
35. Beck, A.; Teboulle, M.: A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 693–696. IEEE (2009)
36. Schmidt, M.; Fung, G.; Rosales, R.: Optimization methods for l1-regularization. University of British Columbia, Technical Report TR-2009, p. 19 (2009)
37. Yang, J.; Zhang, D.; Frangi, A.F.; et al.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **26**(1), 131–137 (2004)
38. Candès, E.J.; Li, X.; Ma, Y.; et al.: Robust principal component analysis? J. ACM (JACM) **58**(3), 11 (2011)