IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 10, 2017, accepted July 17, 2017, date of publication July 21, 2017, date of current version August 8, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2730218

An Improved Kernel Minimum Square Error Classification Algorithm Based on $L_{2,1}$-Norm Regularization

ZHONGHUA LIU[1], SHAN XUE[2], LIN ZHANG[1], JIEXIN PU[1], AND HAIJUN WANG[3]

[1]Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China
[2]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
[3]School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471023, China

Corresponding author: Shan Xue (e-mail: apledxueshan@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U1504610, in part by the Natural Science Foundations of Henan Province under Grant 14A413013, Grant 142102210584, and Grant 18A120002, and in part by the Henan International Cooperation Project under Grant 152102410036.

ABSTRACT The kernel minimum square error classification (KMSEC) algorithm has been widely used in classification problems. It shows a good performance on image data besides the following drawbacks: not sparse in the solutions and sensitive to noises. The latter drawback will result in a decrease in the recognition performance. To this end, we propose an improved (IKMSEC) by using the $L_{2,1}$-norm regularization, which can obtain a sparse representation of nonlinear features to guarantee an efficient classification performance. The comprehensive experiments show the promising results in face recognition and image classification.

INDEX TERMS Minimum square error classification, kernel, $L_{2,1}$-norm, pattern recognition.

I. INTRODUCTION

Image recognition, which is one of the hottest topics in the fields of computer vision and pattern recognition, is receiving much attention and becoming more acknowledged [1]–[5]. Specially, feature extraction and classifier construction are two key processes in designing an image recognition system [6]–[8]. For image recognition, the former researchers have proposed a large number of methods, e.g., linear discriminant analysis (LDA) [9]–[11], principal component analysis (PCA) [12], independent component analysis (ICA) [14], [15], sparse representation-based classifier (SRC) [16]–[20] and minimum squared error algorithm (MSE) [13]. Minimum squared error classification (MSEC), which very easy to be realized by using the sample and its class label as the input and output respectively, gains its popularity in recent years. MSEC is an efficient and simple image recognition method. In addition, MSEC is a special case of LDA when the training samples size approaches infinity. However, due to the inherent linearity, the nonlinear structure of training data in image recognitions can't be represented effectively by above mentioned methods.

The kernel method is firstly applied to support vector machine (SVM) and has achieved a great success [21], [22]. Afterwards, a plenty of methods are proposed for nonlinear data by means of kernel function, e.g., kernel PCA (KPCA) [23], [24], kernel LDA (KLDA) [25], kernel ICA (KICA) [26] and kernel MSEC (KMSEC) [27], [28]. KMSEC maps a low dimensional input to a high dimensional feature space using a nonlinear function, and then accomplishes image recognition. As a matter of fact, KMSEC can be regarded as a feature extraction process based on KPCA, and it is formally equivalent to the least square SVM and the kernel discriminant analysis. When the size of training samples approaches infinite, KMSEC shares the similar idea with the Bayesian discriminant function in the feature space based on the minimum mean square error.

However, the KMSEC is based on Frobenius norm. The main drawback of Frobenius norm is as follows. It can enlarge the noise and error in data. That is to say, it is sensitive to outliers and noises in data, which may decrease the classification accuracy. Resently, the sparsity regularization has been studied to boost the classification performance. Its advantages are listed in the following aspects. 1) It usually leads to more robust models. 2) It avoids overfitting. 3) It may distinguish the most-relevant features. 4) It is able to better use the prior information. Wright *et al.* [29] proposed a classification method based on the sparse representation, named sparse principal component analysis (SPCA), which produces modified principal components with sparse loadings. Qiao *et al.* [31] presented a sparse preserving projection technique, which is proved to be effective in face recognitions. Cai *et al.* [32] proposed a sparse projection algorithm based on graph and showed a good classification performance for document. To improve the classification accuracy in face recognition, a sparse representation algorithm based on kernel method was put forward by Zhu *et al.* [33]. Zhang *et al.* [34] constructed a kernel sparse representation based classifier (KSRC). Based on the fact that the kernel method can effectively process the nonlinear feature, Gao *et al.* [35] presented a kernel sparse representation (KSR) algorithm. Nie *et al.* [36] came up with a robust feature selection algorithm which made joint $L_{2,1}$-norm minimizations on both regularization and loss function. Sun *et al.* [37] put forward an equation constrained $L_{2,1}$-norm minimization model to solve the problem of the multiple measurement vector. Ren *et al.* [38] established an objective model for classification by adding a regularized $L_{2,1}$-norms minimization constraint. The $L_{2,1}$-norm regularized least square regression model was used to make joint feature selection [39]. According to the sparse representation and the projected regularization scheme, the images were separated into a texture part and a cartoon part in [40]. To improve the robustness of the attribute reduction algorithms, Xia et al. [41] proposed an attribute reduction algorithm based on maximum margin projection and $L_{2,1}$-norm regularization. In view of $L_{2,1}$-norm regularization, Hou *et al.* [42] proposed a novel feature selection algorithm, which significantly improved the accuracy in feature selection and the speed in convergence.

In this paper, we propose a robust and efficient classification algorithm by employing the $L_{2,1}$-norm minimization into the KMSEC, which is named IKMSEC. We employ the $L_{2,1}$-norm regularization in our work instead of using the L_2-norm regularization so that the negative impact of the outliers and contaminated samples can be decreased. Moreover, the use of the $L_{2,1}$-norm regularization allows us to select features.

The remainder of the paper is organized as follows. The related work is briefly reviewed in the section II. The proposed IKMSEC based on $L_{2,1}$-norm regularization is presented in section III. Section IV conducts a series of experiments on face recognition databases. The final section concludes the paper.

II. THE RELATED WORK

A. THE MINIMUM SQUARED ERROR CLASSIFICATION (MSEC)

We assume there are C classes for N training images $(x_1, x_2, \cdots, x_N)$ and n training samples for each class, i.e., $N = Cn$. Each training sample is mapped into a vector of values of features. A C-length vector is adopted to stand for the label of a sample. Suppose one sample belongs to the k^{th} class, then the label of this sample is represented as $g_k = [\underbrace{0\ 0\ \cdots\ 0}_{k-1}\ 1\ 0 \cdots\ 0]$. In this way, we acquire the training dataset $X = [x_1, x_2, \cdots, x_N]^T$ and the corresponding label space $G = [g_1^T, g_2^T, \cdots, g_N^T]^T$ for MSEC. MSEC follows the equation:

$$XA = G \tag{1}$$

where A is a transform matrix.

The transform matrix A is generated through $\tilde{A} = (X^T X + \lambda I)^{-1} X^T G$, where λ is a small positive constant and I denotes an identity matrix.

The class label of the test sample y can be obtained by using Eq. (2):

$$g_y = y\tilde{A} \tag{2}$$

We use Eq. (3) to evaluate the Euclidean distances between g_y and each class label g_i, respectively.

$$dt_i = \left\| g_y - g_i \right\|_2, \quad i = 1, 2, \cdots, C \tag{3}$$

If $k = \arg\min_i dt_i$, then the testing sample y is ultimately assigned to the k^{th} class.

B. THE KERNEL MINIMUM SQUARED ERROR CLASSIFICATION (KMSEC)

KMSEC, which mainly deals with the classification issue of nonlinear feature, introduces the nonlinear kernel to the MSEC. Assume that each sample could be mapped from an original low-dimensional feature space to a high-dimensional feature space by a nonlinear mapping function. Given a set of training image samples are denoted as $\tilde{X} = [f(x_1), f(x_2), \cdots, f(x_N)]^T$ in the high-dimensional feature space. KMSEC follows the equation:

$$\tilde{X}\tilde{A} = G \tag{4}$$

where $\tilde{A}$ is a transform matrix, G is same as the G in Eq. (1).

Let $\tilde{A} = [a_1, a_2, \cdots, a_C]$, where $a_i (i = 1, 2, \cdots, C)$ is a vector. According to the correlation theory of kernel function [21], [22], a_i can be expressed as a linear combination of $f(x_j) (j = 1, 2, \cdots, N)$ as follows:

$$\begin{aligned} a_i &= \beta_{i1} f(x_1) + \beta_{i2} f(x_2) + \cdots + \beta_{iN} f(x_N) \\ &= (f(x_1), f(x_2), \cdots, f(x_N)) \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iN} \end{pmatrix} \\ &= (f(x_1), f(x_2), \cdots, f(x_N)) \beta_i \end{aligned} \tag{5}$$

where $\beta_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{iN}]^T$. Then,

$$\begin{aligned} \tilde{X}\tilde{A} &= \tilde{X}[\tilde{X}^T \beta_1, \tilde{X}^T \beta_2, \cdots, \tilde{X}^T \beta_C] \\ &= \tilde{X}\tilde{X}^T [\beta_1, \beta_2, \cdots, \beta_C] \\ &= [f(x_1), f(x_2), \cdots, f(x_N)]^T [f(x_1), f(x_2), \cdots, f(x_N)] \beta \\ &= \begin{bmatrix} f(x_1)^T f(x_1) & f(x_1)^T f(x_2) & \cdots & f(x_1)^T f(x_N) \\ f(x_2)^T f(x_1) & f(x_2)^T f(x_2) & \cdots & f(x_2)^T f(x_N) \\ \vdots & \vdots & \vdots & \vdots \\ f(x_N)^T f(x_1) & f(x_N)^T f(x_2) & \cdots & f(x_N)^T f(x_N) \end{bmatrix} \beta \end{aligned} \tag{6}$$

where $\beta = [\beta_1, \beta_2, \cdots, \beta_C]$. If the kernel function of KMSEC is defined as follows:

$$k(x_i, x_j) = f(x_i)^T f(x_j) \tag{7}$$

Eq. (6) can be rewritten as:

$$\begin{aligned}\tilde{X}\tilde{A} &= \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) \end{bmatrix} \beta \\ &= K\beta \end{aligned} \tag{8}$$

Thus, Eq. (4) can be written as follows:

$$K\beta = G \tag{9}$$

β can be solved by Equation (10)

$$\beta = (K^T K)^{-1} K^T G \tag{10}$$

III. THE IMPROVED KERNEL MINIMUM SQUARED ERROR CLASSIFICATION (IKMSEC) BASED ON $L_{2,1}$-NORM REGULARIZATION

We expect that the transform matrix maintains the sparsity property for feature ranking. To this end, a $L_{2,1}$-norm regularization is applied to minimize β, denoted as $\|\beta\|_{2,1}$. The objective function of the proposed IKMSEC is formulated as:

$$\min_{\beta} L_\beta = \min_{\beta}(\|K\beta - G\| + \gamma \|\beta\|_{2,1}) \tag{11}$$

where the $L_{2,1}$-norm of the matrix β is defined as follows:

$$\|\beta\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} \beta_{ij}^2} = \sum_{i=1}^{n} \left\|\beta^i\right\|_2 \tag{12}$$

where β^i is the i^{th} row of matrix β. Then, we have

$$\|\beta\|_{2,1} = \beta^T U \beta \tag{13}$$

where $U_{i,i} = \frac{1}{2\|\beta^i\|_2}$, if $i = j$, $U_{ij} = 0$.

The objective function of IKMSEC is optimized as:

$$\begin{aligned} \mathrm{L}(\beta) &= ||K\beta - G||_2^2 + \gamma ||\beta||_{21} \\ &= trace(\beta^T K^T K\beta - 2\beta^T K^T G + G^T G) + \gamma\, trace(\beta^T U\beta) \\ &= trace(\beta^T(K^T K + \gamma U)\beta - 2\beta^T K^T G + G^T G) \end{aligned} \tag{14}$$

The solution of Eq. (14) can be generated through solving the least squares below.

$$(K^T K + \gamma U)\beta = K^T G \tag{15}$$

Since Eq. (15) is solved by iterations on $U_{ii} = \frac{1}{2||\beta^i||}$, we initialize U_0 as an identity matrix and define U_{t+1} in the iterations as:

$$U_{t+1} = \begin{bmatrix} \frac{1}{2\left\|\beta_t^1\right\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\left\|\beta_t^m\right\|_2} \end{bmatrix} \tag{16}$$

The proposed algorithm is robust, so it can well overcome the uncertainty problem in face images. Namely, different images of the same face tend to vary significantly. At the same time, the proposed algorithm is less sensitive to outliers and noise. In addition, the proposed algorithm can get the sparse solution.

The proposed IKMSEC algorithm is presented in detail as follows.

Step 1. Initialize U in Eq. (15) as an identity matrix, denoted as U_0.

Step 2. Generate the iterations on U_{t+1} in Eq. (16) and update β_{t+1} by Eq. (15).

Step 3. Check the convergence of Eq. (14): output the optimal matrix $\beta^* = \beta_t$, if convergence; and go to Step 2, otherwise.

IV. EXPERIMENTS AND RESULTS

In this subsection, we implement comprehensive experiments to assess the effectiveness and the robustness of the proposed IKMSEC on databases of Olivetti Research Laboratory (ORL) [43], Yale, FERET [44], Georgia Tech and AR face [45]. We apply the Gaussian kernel function ($k(x, y) = \exp(-\|x - y\|^2/t)$) and the Polynomial kernel function ($k(x, y) = (x^T \times y)^d$) for KMSEC and IKMSEC. The optimal kernel parameters t and d are used in the experiments.

A. EXPERIMENT ON ORL FACE DATABASE

In the ORL face image database, there are 400 samples of images from 40 objects, that each object has 10 samples. The images include different expressions (non-smiling and smiling, closed and open eyes), different facial details (no glasses and glasses), and different angles (a tolerance for the rotations and tilting up to 20 degrees). In addition, the samples of images are captured at different times. The size of the image samples are cropped to 56×46 pixels. Fig. 1 illustrates a set of samples of one object.

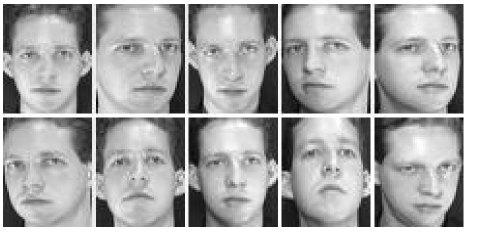


FIGURE 1. Image samples of one object in the ORL database.

In this experiment, the first l samples of every object (l varies from 1 to 9) are chosen as the training set, and the rest samples are used as the testing set. The parameter of t in the Gaussian kernel function is set as 1 for KMSEC and IKMSEC. The parameter of d in the Polynomial kernel function is set as 8 for KMSEC and IKMSEC. Fig. 2 shows the face recognition results.

In Fig. 2, there are three main findings. To begin with, whatever kernel function we choose, the proposed IKMSEC

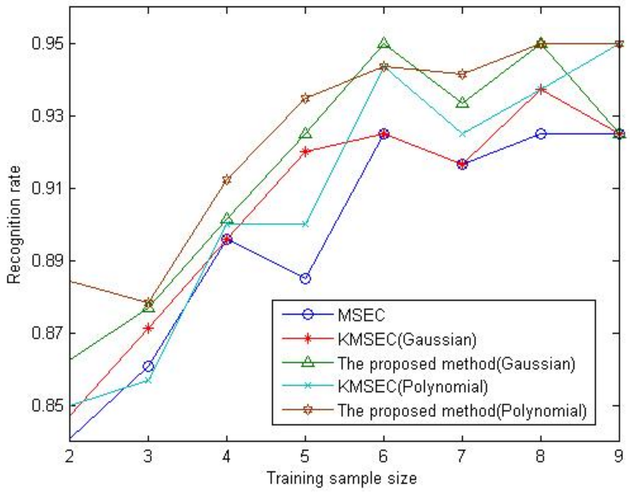


FIGURE 2. **The recognition accuracies of several different methods with different numbers of training samples on the ORL image database.**

algorithm always outperforms KMSEC irrespective of the variation of the training sample size. Then, the performance of KMSEC is better than that of MSEC. Last, from the two points above, it is obvious to tell that the kernel method promotes the classification performances. At the same time, the proposed IKMSEC is more robust than the original KMSEC.

B. EXPERIMENT ON YALE FACE DATABASE

The Yale face image database has 15 objects and a total of 165 images, i.e., each object contains 11 image samples with different illuminations and facial expressions. In this experiment, all image samples are cropped to 50×40 pixels. Fig. 3 presents the samples of one object.

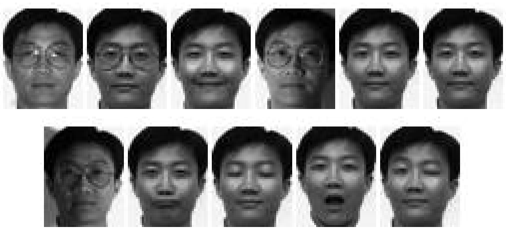


FIGURE 3. **Image samples of one object in the Yale database.**

In the experiment, the number of training samples of each object varies from 1 to 10, and the rest image samples are used as the test samples. The parameter of the Gaussian kernel function is set as $t = 2$ for KMSEC and IKMSEC. The parameter d of the Polynomial kernel function is set as 0.8 for KMSEC and IKMSEC. Fig. 4 shows the face recognition accuracies of all methods with different numbers of training samples.

It can be seen from Fig. 4 that the proposed IKMSEC method still performs the best regardless of the number of training samples. Because the face database is relatively small, all the methods have achieved good recognition performance. When the training sample size is larger than 3,

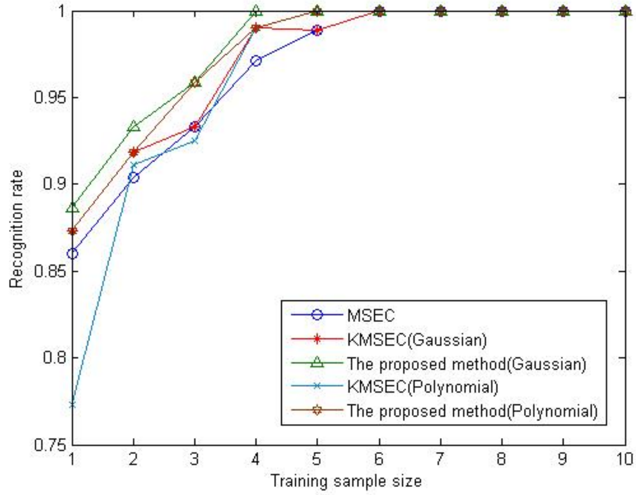


FIGURE 4. **The recognition accuracies of sereral methods with different numbers of training samples on the Yale database.**

IKMSEC based on the Gaussian kernel function consistently achieves 100% in classification accuracy. IKMSEC based on the Polynomial kernel function achieves 100% in classification accuracy when the training sample size is larger than 4. Other methods achieves 100% in classification accuracy when the training sample size is larger than 6.

C. EXPERIMENT ON FERET FACE DATABASE

There are 1,565 objects and a total of 13,539 image samples in the FERET database. The image samples vary in illuminations, ages, sizes, facial expressions, and poses. We select 1400 images from 200 objects (each object includes 7 image samples) in this experiment. Each sample is resized to 40×40 pixels. Fig. 5 demonstrates the samples of one object.

FIGURE 5. **Image samples of one object in the FERET database.**

In this experiment, the first l image samples (l varies from 1 to 6) of every object are selected as the training samples and the rest of images are selected as the test samples. The parameter of the Gaussian kernel function is set as $t = 1$ for KMSEC and IKMSEC. The parameter d of the Polynomial kernel function is set as 10 for KMSEC and is set as 20 for IKMSEC. Fig. 6 illustrates the recognition accuracies of different methods based on different training sample sizes.

In Fig. 6, we find that our proposed IKMSEC method outperforms KMSEC and MSEC in the classification task irrespective of the number of the training samples. Besides, the accuracy of KMSEC method outperforms that of MSEC only when the training sample size is 3. The recognition rates of KMSEC and the proposed IKMSEC with Gaussian kernel function are respectively 67.75% and 74%. The recognition

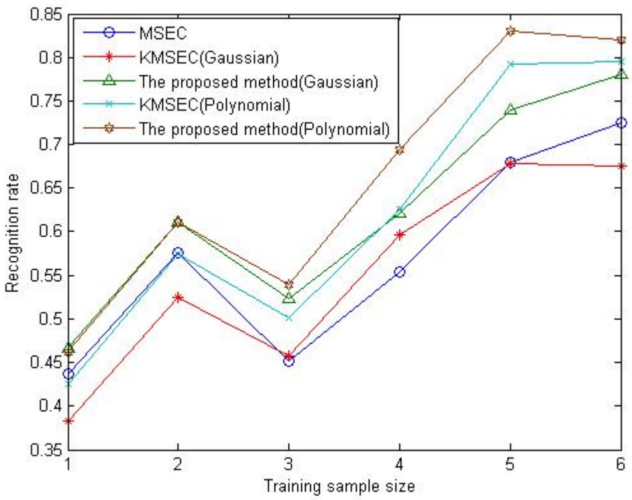


FIGURE 6. The recognition accuracies of several methods with different numbers of training samples on the FERET database.

rates of KMSEC and the proposed IKMSEC with Polynomial kernel function are respectively 67.25% and 72.75%.

D. EXPERIMENT ON GEORGIA TECH IMAGE DATABASE

Georgia Tech image database includes 50 objects taken in two or three sessions by GIT (Georgia Institute of Technology). The 15 color JPEG images are captured at the resolution of 640×480 pixels for each object in this database. These samples are tilted or fronted with various scales, expressions, and illuminations. All samples are resized to 40×30 pixels in this experiment. Fig. 7 illustrates the samples of one object.

FIGURE 7. Image samples of one object in the Georgia Tech face database.

For this experiment, all the images are firstly converted into grayscale images. We select the first 1, 2, up to 14 image samples of every object to be the training set and the rest as test set. The parameter of the Gaussian kernel function is set as $t = 1$ for KMSEC and $t = 0.5$ for IKMSEC. The parameter d of the Polynomial kernel function is set as 10 for KMSEC and 15 for IKMSEC. Fig. 8 illustrates the recognition accuracies with different training sample sizes.

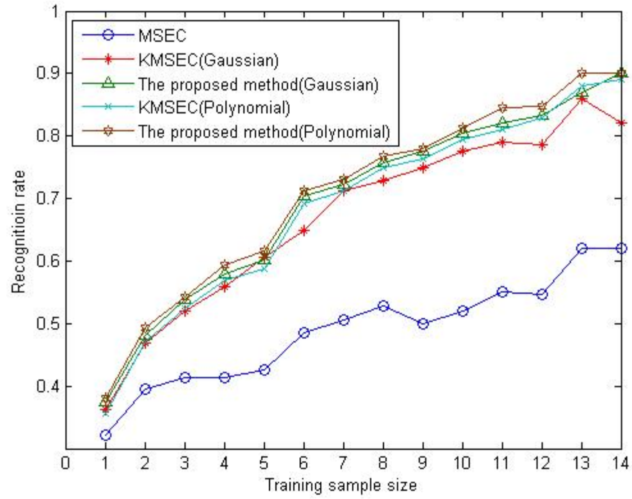


FIGURE 8. The recognition accuracies of several methods with different numbers of training samples on the GT database.

Fig. 8 shows that the proposed IKMSEC method still performs the best irrespective of the number of training samples and the recognition performance of the proposed method and KMSEC is much better than that of MSEC. However, the performance of the proposed method and KMSEC is relatively close. This is because the pose of face images in GT database varies greatly, and the proposed method is robust to noise and outliers.

E. EXPERIMENT ON AR IMAGE DATABASE

There are over 400 images of color faces from 126 objectives in the AR face database. Under different lighting conditions, each object has 26 frontal views of faces with occlusions and various facial expressions. The AR image database is taken in two sessions and each session (14 days) includes 13 color images. In this experiment, we select 14 non-occluded face images (with 7 samples of every session for every object), and the image samples are transformed to grayscale. The size of each image is converted to 50×40 pixels. The image samples of one object are shown in Fig. 9.

FIGURE 9. Image samples of one object with two sessions in the AR database.

In this experiment, we set the first 1, 2, up to 7 image samples of each object from the first session as the training set, and all the 7 images of each object from the second session are used as the test set. The parameter in the Gaussian kernel function is set as $t = 0.2$ for KMSEC and $t = 0.3$ for IKMSEC. The parameter in the Polynomial kernel

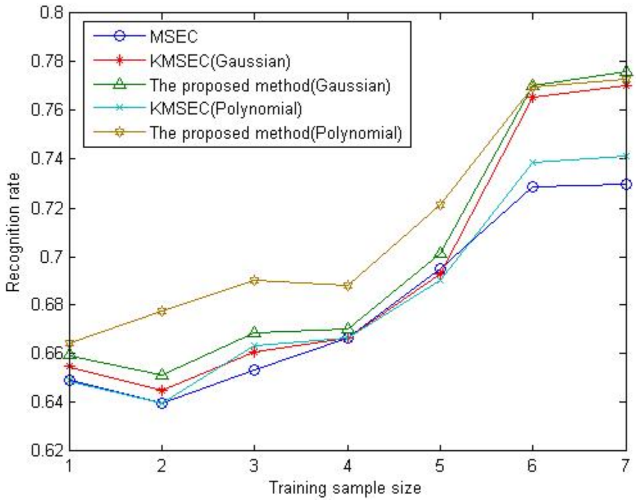


FIGURE 10. The recognition accuracies of several methods with different numbers of training samples on the AR database.

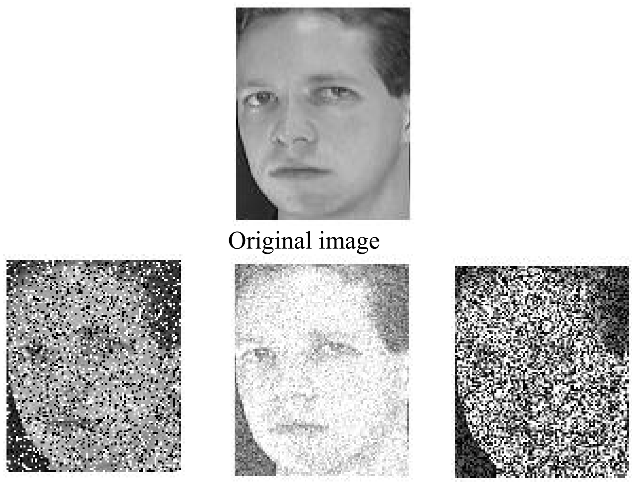


FIGURE 11. One image and its corresponding corrupted sample images.

function is set as $d = 0.4$ for KMSEC and IKMSEC. The experiment results are illustrated in Fig. 10. It can be seen from Figure 10 that the proposed IKMSEC shows a better recognition performance than KMSEC regardless of the kernels. The performance of the proposed IKMSEC with Polynomial kernel function is more than that of the proposed IKMSEC with Gaussian kernel function.

F. EXPERIMENT ON NOISY ORL DATABASE

For the sake of validating the robustness of the proposed IKMSEC for noise, we implement some experiments in this subsection. The first l image samples each object (l varies from 1 to 9) are given to act as the training set, and the rest as the test set. We give the following assumptions: the training samples are not corrupted, and the test samples are corrupted by three types of noise, i.e., Gaussian noise, Speckle noise and Salt & pepper noise. Fig. 11 illustrates one image and its corresponding corrupted images. We use the Polynomial kernel function and the Gaussian kernel function for KMSEC and IKMSEC. The recognition results are listed in Fig. 12 (based on the Gaussian kernel function) and Fig. 13 (based on the Polynomial kernel function). It can be known from Fig. 12 and Fig. 13 that the recognition performance of our IKMSEC method is much better than that of the original KMSEC method regardless of any kind of noises. In this experiment, the experiment results prove that our IKMSEC method is robust to noises.

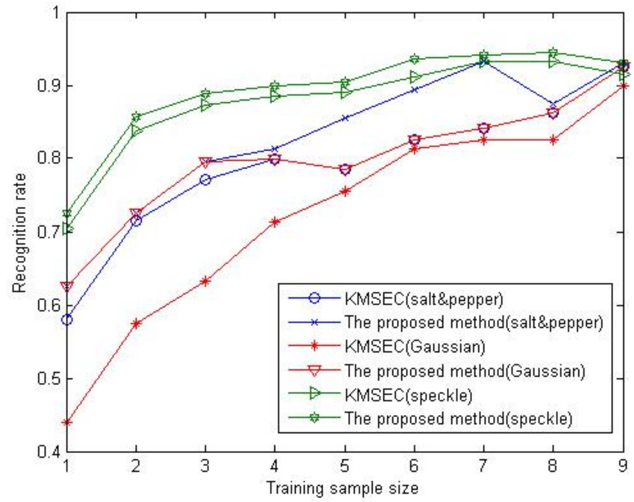


FIGURE 12. The recognition rates of several methods with Gaussian kernel function versus the variation of the training sample size on the corrupted ORL database.

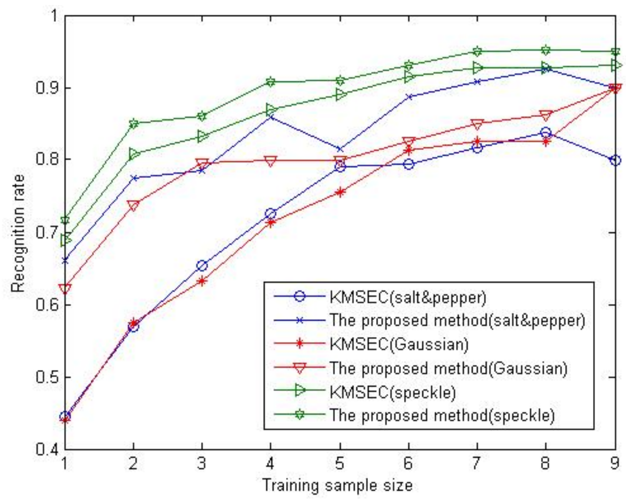


FIGURE 13. The recognition rates of several methods with Polynomial kernel function versus the variation of the training sample size on the corrupted ORL database.

V. CONCLUSION

In this paper, we present an improved kernel minimum square error classification algorithm based on $L_{2,1}$-norm, which is a kind of sparse coding techniques that applied in a high dimensional feature space and mapped by an implicit nonlinear function. The proposed method is robust, and it can overcome the uncertainty problem in face images. The proposed method has been successfully used in solving face classification problems. The experiment results on five benchmark face databases show that our algorithm is very efficient in image classification tasks, and outperforms KMSEC methods and MSEC methods.

REFERENCES

[1] J. Wu, Z. Hong, S. Pan, X. Zhu, Z. Cai, and C. Zhang, "Multi-graph-view learning for graph classification," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2014, pp. 590–599.

[2] Y. Xu, X. Li, J. Yang, Z. Lai, and D. Zhang, "Integrating conventional and inverse representation for face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1738–1746, Oct. 2014.

[3] J. Zheng, P. Yang, G. Shen, W. Wang, and S. Chen, "Iterative re-constrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, May 2017, doi: 10.1109/TIP.2017.2681841.

[4] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.

[5] J. Wu, S. Pan, X. Zhu, C. Zhang, and X. Wu, "Positive and unlabeled multi-graph learning," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 818–829, Apr. 2017.

[6] J. Wu, S. Pan, X. Zhu, C. Zhang, and P. Yu, "Multiple structure-view learning for graph classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2703832.

[7] Y. Xu, X. Zhu, Z. Li, G. Liu, Y. Lu, and H. Liu, "Using the original and 'symmetrical face' training samples to perform representation based two-step face recognition," *Pattern Recognit.*, vol. 46, no. 4, pp. 1151–1158, 2013.

[8] J. Wu, S. Pan, X. Zhu, P. Zhang, and C. Zhang, "SODE: Self-adaptive one-dependence estimators for classification," *Pattern Recognit.*, vol. 51, pp. 358–377, Mar. 2016.

[9] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[10] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 14, no. 8, pp. 1724–1733, 1997.

[11] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[12] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 8–21, Jan. 2013.

[13] A. Hyvärinen, "Survey on independent component analysis," *Neural Comput. Surv.*, vol. 2, no. 4, pp. 94–128, 1999.

[14] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski, "Independent component representations for face recognition," *Proc. SPIE*, vol. 3299, pp. 528–539, Jan. 1998.

[15] C. H. Park and H. Park, "A relationship between linear discriminant analysis and the generalized minimum squared error solution," *SIAM J. Matrix Anal. Appl.*, vol. 27, no. 2, pp. 474–492, 2005.

[16] Y. Xu, Z. Fan, M. Qiu, D. Zhang, and J.-Y. Yang, "A sparse representation method of bimodal biometrics and palmprint recognition experiments," *Neurocomputing*, vol. 103, pp. 164–171, Mar. 2013.

[17] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Kernel sparse representation for image classification and face recognition," in *Computer Vision—ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer-Verlag, 2010, pp. 1–14.

[18] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, "A two-phase test sample sparse representation method for use with face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1255–1262, Sep. 2011.

[19] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L_1-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, 2012.

[20] L. Zhang, S. Chen, and L. Qiao, "Graph optimization for dimensionality reduction with sparsity constraints," *Pattern Recognit.*, vol. 45, no. 3, pp. 1205–1210, 2012.

[21] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[22] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[23] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[24] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. ICCANN,* 1997, pp. 583–588.

[25] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. R. Müllers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Process. Soc. Workshop Neural Netw. Signal Process.*, vol. 9. Aug. 1999, pp. 41–48.

[26] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, Jan. 2002.

[27] M. L. Visinsky, J. R. Cavallaro, and I. D. Walker, "A dynamic fault tolerance framework for remote robots," *IEEE Trans. Robot. Autom.*, vol. 11, no. 4, pp. 477–490, Aug. 1995.

[28] Y. Xu, "A new kernel MSE algorithm for constructing efficient classification procedure," *Int. J. Innov. Comput., Inf. Control*, vol. 5, no. 8, pp. 2439–2447, 2009.

[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[30] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, 2006.

[31] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.

[32] D. Cai, X. He, and J. Han, "Sparse projections over graph," in *Proc. AAAI*, 2008, pp. 610–615.

[33] N. Zhua, T. Tanga, S. Tang, D. Tang, and F. Yu, "A sparse representation method based on kernel and virtual samples for face recognition," *Optik-Int. J. Light Electron Opt.*, vol. 124, no. 23, pp. 6236–6241, 2013.

[34] L. Zhang *et al.*, "Kernel sparse representation-based classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, Apr. 2012.

[35] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Sparse representation with kernels," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 423–434, Feb. 2013.

[36] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. NIPS*, 2010, pp. 1813–1821.

[37] L. Sun, J. Liu, J. Chen, and J. Ye, "Efficient recovery of jointly sparse vectors," in *Proc. NIPS*, 2009, pp. 1812–1820.

[38] C.-X. Ren, D.-Q. Dai, and H. Yan, "Robust classification using $\ell_{2,1}$-norm based regression model," *Pattern Recognit.*, vol. 45, no. 7, pp. 2708–2718, 2012.

[39] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," in *Proc. UAI*, 2009, pp. 339–348.

[40] L. Jiang, H. Yin, and X. Feng, "Image decomposition based on sparse representations and a projected regularization method," *J. Xidian Univ.*, vol. 34, no. 5, pp. 800–804, 2007.

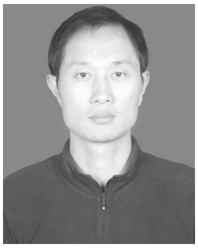
[41] J. M. Xia and J. A. Yang, "A novel attribute reduction algorithm based on maximum margin projection and $\ell_{2,1}$ norm regularization," *Control Decision*, vol. 28, no. 10, pp. 1485–1490, 2013.

[42] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. IJCAI*, 2011, pp. 1324–1329.

[43] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. WACV*, Dec. 1994, pp. 138–142.

[44] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[45] A. Martinez and R. Benavente, "The AR face database," Centre Visió Comput., Purdue Univ., West Lafayette, IN, USA, Tech. Rep., 1998.



ZHONGHUA LIU received the B.S. degree in computer engineering from the First Aeronautical Institute of the Air Force, the M.S. degree in computer software and theory from Xihua University, and the PhD degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology in 1998, 2005, and 2011, respectively. He is an Associate Professor with the Information Engineering College, Henan University of Science and Technology. His current research interests include pattern recognition, face recognition, image processing, and scene matching.

SHAN XUE is currently pursuing the dual Ph.D. degrees with the Center for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney and with the School of Management, Shanghai University. Her research focuses on artificial intelligence, machine learning, and information management.

LIN ZHANG is currently pursuing the M.S. degree in the Information Engineering College, Henan University of Science and Technology, China. Her current research interests include pattern recognition and machine learning.

JIEXIN PU is a Professor with the Information Engineering College, Henan University of Science and Technology. His research interests include computer application technology, pattern recognition and computer network, and information security.

HAIJUN WANG received the B.Eng. degree in computer software and the M.Eng. degree in computer software and theory from Hunan University, Changsha, China, in 2002 and 2007, respectively., and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2016. Since 2002, he has been with the School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang, China. His research interests include machine learning and computer vision.