Journal of Information Hiding and Multimedia Signal Processing ©2014 ISSN 2073-4212
Ubiquitous International Volume **5**, Number **1**, January **2014**

Optimizing Matrix Mapping with Data Dependent Kernel for Image Classification

De-Long Zhang, Jiaqing Qiao, Jun-Bao Li, and Li-Yan Qiao
Department of Automatic Test and Control
Harbin Institute of Technology, China
junbaolihit@gmail.com

Shu-Chuan Chu, and John F. Roddick
School of Computer Science, Engineering and Mathematics
Flinders University
GPO Box 2100, Adelaide SA 5001
scchu@bit.kuas.edu.tw

Received July, 2013; revised September, 2013

Abstract. *Kernel based nonlinear feature extraction is feasible to extract the feature of image for classification.The current kernel-based method endures two problems: 1) kernel-based method is to use the data vector through transforming the image matrix into vector, which will cause the store and computing burden; 2) the parameter of kernel function has the heavy influences on kernel based learning method. In order to solve the two problems, we present the method of optimizing matrix mapping with data dependent kernel for feature extraction of the image for classification. The method implements the algorithm without transforming the matrix to vector, and it adaptively optimizes the parameter of kernel for nonlinear mapping. The comprehensive experiments are implemented evaluate the performance of the algorithms.*

Key words: Image classification, feature extraction, matrix mapping, data dependent kernel.

1. **Introduction.** Face recognition has received more attention owing to its potential applications in law enforcement, information security and surveillance, smart cards, and so on [1], [2], [3]. Variations of illumination degrade the recognition performance of most current technologies for face recognition in both academic face recognition algorithms and commercial systems. The low-dimensional feature representation with high discriminatory power is very important for facial feature extraction, such as principal component analysis (PCA) and linear discriminant analysis (LDA) [4], [5], [6]. Recently, researchers applied kernel machine techniques to solve the nonlinear problem successfully [7], [8], [9], and accordingly some kernel based methods are developed for face recognition [10]-[14]. The method achieves the excellent performances on image denoising, image debluring, image restoration, and so on, and they are the basic operation in the image processing [20]-[23]. Among the mentioned face recognition methods, kernel-based nonlinear feature extraction methods perform better than other linear ones, which indicates kernel-based method is promising for face recognition. Extracting the Gabor features of face image is feasible to deal with the variations in illumination and facial expression. The distribution of feature information mostly converges in the low frequency phase, but the overfull high frequency information influences representation of the discriminant features for face

recognition. Kernel functions and the corresponding parameters have high influence the recognition performance straightly, and the procedural parameters are chosen through cross-validation method, so how to choose them automatically should be researched in the future work. But current kernel based facial feature extraction methods face the following problems. 1) Current face recognition methods are based on image or video, while the current popular kernels need format of the input data is a vector. Thus kernel based facial feature extraction causes the large storage requirements and the large computational effort for transforming images to vectors owing to its viewing images as vectors. 2) Different kernels can cause the different RKHS in which the data has different class discrimination, so the selection of kernels will influence the recognition performance of the kernel based methods. And the inappropriate selection of kernels will decrease the performance. But unfortunately the geometrical structures of the data in the feature space will not be changeable when we only change the parameter of the kernel.

In this paper, we present the method of optimizing matrix mapping with data dependent kernel for feature extraction of the image for classification. Firstly we create a novel matrix norm based Gaussian kernel, which views images as matrices for facial feature extraction, as the basic kernel of data-dependent kernel, while the data-dependent kernel can change the geometrical structures of the data with different expansion coefficients. And then we apply the maximum margin criterion to solve the adaptive expansion coefficients of ADM-Gaussian kernel which leads the largest class discrimination in the feature space.

2. **Optimizing matrix mapping with data-dependent kernel.** In this section, firstly we introduce the data-dependent kernel based on vectors and then we extend it to the version of matrices. Secondly we introduce the theoretical analysis of matrix norm based Gaussian kernel, and finally we apply the maximum margin criterion to seek the adaptive expansion coefficients of the data-dependent matrix norm based Gaussian kernel.

Data-dependent kernel with a general geometrical structure is applied to create a new kernel in this paper. Given a basic kernel $k_b(x,y)$, its data-dependent kernel $k_d\,(x,y)$ can be defined as follows.

$$k_d\,(x,y) = f(x)f(y)k_b(x,y) \tag{1}$$

where $f(x)$ is a positive real valued function x, which is defined as follows.

$$f(x) = b_0 + \sum_{n=1}^{N_{XV}} b_n e(x, \widetilde{x}_n) \tag{2}$$

In the previous work in [15], Amari and Wu expanded the spatial resolution in the margin of a SVM by using $f(x) = \sum_{i\in SV} a_i e^{-\delta|x-\widetilde{x}_i|^2}$, where $\widetilde{x}_i$ is the ith support vector, SV is a set of support vector, a_i is a positive number representing the contribution of $\widetilde{x}_i$, and δ is a free parameter.

We extend it to the matrix version and propose the Adaptive Data-dependent Matrix Norm Based Gaussian Kernel (ADM-Gaussian kernel) as follows. Supposed that $k_b(X,Y)$ is so-called matrix norm based Gaussian kernel (M-Gaussian kernel) as the basic kernel and $k_d(X,Y)$ is a data-dependent matrix norm based Gaussian kernel. Then data-dependent matrix norm based Gaussian kernel is defined as follows.

$$k_d\,(X,Y) = f(X)f(Y)k_b(X,Y) \tag{3}$$

where $f(X)$ is a positive real valued function X,

$$f(X) = b_0 + \sum_{n=1}^{N_{XV}} b_n e(X, \widetilde{X}_n) \tag{4}$$

where $e(X, \widetilde{X}_n)(1 \le n \le N_{XM})$ is defined as follows.

$$e(X, \widetilde{X}_n) = \exp\left(-\delta \sum_{j=1}^{N} \left(\sum_{i=1}^{M} (x_{ij} - \widetilde{x_{ij}})^2\right)^{1/2}\right) \tag{5}$$

where $\widetilde{x_{ij}}(i = 1, 2, \dots, M, j = 1, 2, \dots, N)$ is the elements of matrix $\widetilde{X_n}(n = 1, 2, \dots, N_{XM})$, and δ is a free parameter, and $\widetilde{X}_n$, $1 \le n \le N_{XM}$, are called the "expansion matrices (XMs)" in this paper, N_{XM} is the number of XMs, and $b_i \in R$ is the "expansion coefficient" associated with $\widetilde{X}_n$. The $\widetilde{X}_n$, $1 \le n \le N_{XM}$, for its vector version, have different notations in the different kernel learning algorithms.

Given n samples X_p^q $(X_p^q \in \mathrm{R}^{M \times N})$, $(p = 1, 2, \dots, L,\ q = 1, 2, \dots, n_p)$ where n_p denotes the number of samples in the pth class and L denote the number of the classes. M-Gaussian kernel $k_b(X, Y)$ is defined as follows.

$$k\left(X, Y\right) = \exp\left(-\frac{\sum\limits_{j=1}^{N}\left(\sum\limits_{i=1}^{M}\left(x_{ij} - y_{ij}\right)^2\right)^{1/2}}{2\sigma^2}\right) \quad (\sigma > 0) \tag{6}$$

where $X = [x_{ij}]_{i=1,2,\dots,M; j=1,2,\dots,N}$ and $Y = [y_{ij}]_{i=1,2,\dots,M; j=1,2,\dots,N}$ denote two sample matrices. Now we want to prove that $k\left(X, Y\right) = e^{-\frac{\sum_{j=1}^{N}\left(\sum_{i=1}^{M}\left(x_{ij}-y_{ij}\right)^2\right)^{1/2}}{2\sigma^2}}$ is a kernel function. Kernel function can be defined in various ways. In most cases, however, kernel means a function whose value only depends on a distance between the input data, which may be vectors.

It is a sufficient and necessary condition for a symmetric function to be a kernel function that its Gram matrix is positive semidefinite [16]. Given a finite data set $X = \{x_1, x_2, \dots, x_N\}$ in the input space and a function $k\left(\cdot, \cdot\right)$, the $N \times N$ matrix K with elements $K_{ij} = k\left(x_i, x_j\right)$ is called Gram matrix of $k\left(\cdot, \cdot\right)$ with respect to $x_1, x_2, \dots, x_N$. And it is easy to know that $k\left(X, Y\right) = e^{-\frac{\sum_{j=1}^{N}\left(\sum_{i=1}^{M}\left(x_{ij}-y_{ij}\right)^2\right)^{1/2}}{2\sigma^2}}$ is a symmetric function. The matrix K, which is derived from the $k\left(X, Y\right)$, is positive definite. While it is easy to know that $F(X) = \sum\limits_{j=1}^{N}\left(\sum\limits_{i=1}^{M} x_{ij}^2\right)^{\frac{1}{2}}, (X = [x_{ij}]_{i=1,2,\dots,M; j=1,2,\dots,N})$ is a matrix norm. $k\left(X, Y\right) = e^{-\frac{\sum_{j=1}^{N}\left(\sum_{i=1}^{M}\left(x_{ij}-y_{ij}\right)^2\right)^{1/2}}{2\sigma^2}}$ is derived form the matrix norm, so we can call it a matrix norm based Gaussian kernel. Gaussian kernel denotes the distribution of similarity between two vectors. Similarly M-Gaussian kernel also denotes the distribution of similarity between two matrices. M-Gaussian kernel views an image as a matrix, which enhances the computation efficiency without influencing the performance of kernel-based method.

Our goal is to seek the optimal expansion coefficients for the data-dependent matrix norm based Gaussian kernel (ADM-Gaussian kernel), and the ADM-Gaussian kernel is adaptive to the input data in the feature space. The data in the feature space have the largest class discrimination with the ADM-Gaussian kernel.

Firstly our goal is select the free parameter δ, and the expansion matrices $\widetilde{X}_n$, $1 \leq n \leq N_{XM}$. In this paper, we select the mean of the class as the expansion matrix. That is, $N_{XM} = L$. Let $\overline{X}_n$ denotes the mean of the *nth* class, then

$$e(X, \widetilde{X}_n) = e(X, \overline{X}_n) = \exp\left(-\delta \sum_{j=1}^{N} \left(\sum_{i=1}^{M} (x_{ij} - \overline{x_{ij}})^2 \right)^{1/2} \right) \tag{7}$$

where $\overline{x_{ij}}(i = 1, 2, \ldots, Mj = 1, 2, \ldots, N)$ is the elements of matrix $\overline{X}_n$ $(n = 1, 2, \ldots, L)$.

After we select the expansion vectors and the free parameter, our goal is to find the expansion coefficients varied with the input data to optimize the kernel. According to the equation (2), given one free parameter δ and the expansion vectors $\{\widetilde{x}_i\}_{i=1,2\ldots,N_{XVs}}$, we create a matrix as follows.

$$E = \begin{bmatrix} 1 & e(X_1, \widetilde{X}_1) & \cdots & e(X_1, \widetilde{X}_{N_{XMs}}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e(X_M, \widetilde{X}_1) & \cdots & e(X_M, \widetilde{X}_{N_{XMs}}) \end{bmatrix} \tag{8}$$

Let $\beta = [b_0, b_1, b_2, \ldots, b_{N_{XM}}]^T$ and $\Lambda = diag(f(X_1), f(X_2), \ldots, f(X_n))$, and according to the equation (2), we obtain

$$\Lambda 1_n = E\beta \tag{9}$$

where 1_n is a n-dimensional vector whose entries equal to unity.

Now our goal is to create a constrained optimization function to seek an optimal expansion coefficient vector β In this paper, we apply the maximum margin criterion to solve the expansion coefficients. We maximize the class discrimination in the high-dimensional feature space by maximizing the average margin between different classes which is widely used as maximum margin criterion for feature extraction [17].

So the problem of solving the constrained optimization function is transformed to the problem of solving eigenvalue equation. We can obtain the optimal expansion coefficient vector β^* that is, the eigenvector of $E^T ME$ corresponding to the largest eigenvalue. It is easy to see that the data-dependent kernel with β^* is adaptive to the input data, which leads to the best class discrimination in feature space for given input data.

The procedure of creating the ADM-Gaussian kernel can be described as follows.

Step 1. Compute the basic M-Gaussian kernel matrix $K_b = [k_b(X_i, X_j)]_{n \times n}$with the formulation (6).

Step 2. Compute the matrix E and M with the formulation (8) and the proposition 2.

Step 3 Obtain the adaptive expansion coefficients vector β^* by solving equation (22).

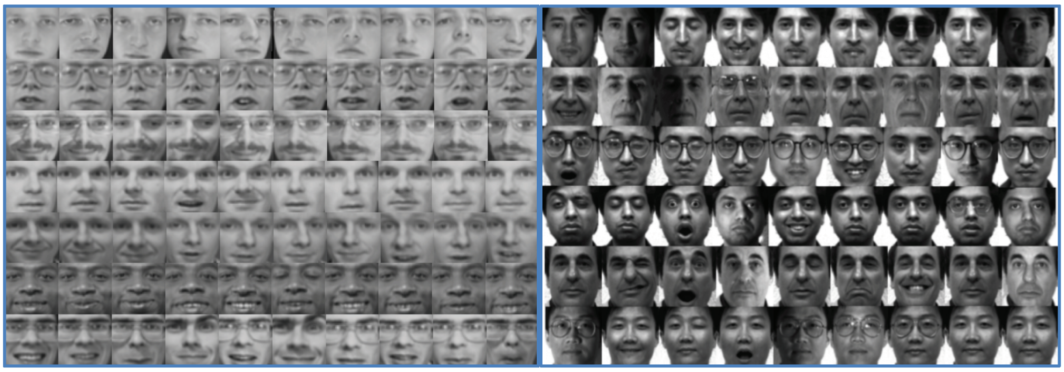
Step 4. Calculate the ADM-Gaussian kernel matrix with the $k_d(X, Y)$ with the optimal expansion coefficients vector β^*

3. **Experimental results.** We implement KPCA with Gaussian kernel, M-Gaussian kernel, and ADM-Gaussian kernel on the two face databases, i.e., ORL face database [19] and Yale face database [1]. We carry out the experiments with two parts as follows. 1) Model selection: selecting the optimal parameters of three Gaussian kernels and the free parameter of the data-dependent kernel for Gaussian kernel, M-Gaussian kernel, and ADM-Gaussian kernel. 2) Performance evaluation: comparing the recognition performance of KPCA with Gaussian kernel, M-Gaussian kernel, and ADM-Gaussian kernel.

In our experiments, we implement our algorithm in the two face databases, ORL face database[19] and Yale face database[1]. The ORL face database, developed at the Olivetti

Research Laboratory, Cambridge, U.K., is composed of 400 grayscale images with 10 images for each of 40 individuals. The variations of the images are across pose, time and facial expression. The Yale face database was constructed at the Yale Center for Computational Vision and Control. It contains 165 grayscale images of 15 individuals. These images are taken under different lighting condition (left-light, center-light, and right-light), and different facial expression (normal, happy, sad, sleepy, surprised, and wink), and with/without glasses.

In our experiments, to reduce computation complexity, we resize the original ORL face images sized 112×92 pixels with a 256 gray scale to 48×48 pixels, and some examples are shown in Fig. 1a. We randomly select 5 images from each subject, 200 images in total for training, and the rest 200 images are used to test the performance. Similarly, the images from Yale databases are cropped to the size of 100×100 pixels, and some examples are shown in Fig. 1b. Randomly selected 5 images per person are selected as the training samples, while the rest 5 images per person are used to test the performance.



(a) (b)

FIGURE 1. Example face images of face databases used in our experiments. (a) Example cropped face images from the ORL face database in our experiments (cropped to the size of 48×48 to extract the facial region). (b) Example cropped face images from the Yale face database in our experiments (cropped to the size of 100×100 to extract the facial region).

In this section, our goal is to select the kernel parameters of three Gaussian kernels and the free parameter of the data-dependent kernel for Gaussian kernel, M-Gaussian kernel, and ADM-Gaussian kernel. We select the following parameters for selection of Gaussian kernel and the free parameter for M-Gaussian kernel, $\sigma^2 = 1 \times 10^5$, $\sigma^2 = 1 \times 10^6$, $\sigma^2 = 1 \times 10^7$ and $\sigma^2 = 1 \times 10^8$ for M-Gaussian kernel parameter, the free parameter of the data-dependent kernel, $\delta = 1 \times 10^5$, $\delta = 1 \times 10^6$, $\delta = 1 \times 10^7$, $\delta = 1 \times 10^8$, $\delta = 1 \times 10^9$, and $\delta = 1 \times 10^{10}$. Moreover, the dimension of the feature vector is set to 140 for ORL face database, and 70 for Yale face base. From the experimental results, we find that the higher recognition rate is can obtained under the following parameters, $\sigma^2 = 1 \times 10^8$ and $\delta = 1 \times 10^5$ for ORL face database, and $\sigma^2 = 1 \times 10^8$ and $\delta = 1 \times 10^7$ for Yale face database. After we select the parameters for AMD-Gaussian kernel, we select the Gaussian parameter for Gaussian kernel and M-Gaussian kernel. $\sigma^2 = 1 \times 10^5$, $\sigma^2 = 1 \times 10^6$, $\sigma^2 = 1 \times 10^7$ and $\sigma^2 = 1 \times 10^8$ is selected to test the performance. From the experiments, we find that, on ORL face database, $\sigma^2 = 1 \times 10^6$ is selected for Gaussian, and $\sigma^2 = 1 \times 10^5$ is selected for M-Gaussian kernel. And $\sigma^2 = 1 \times 10^8$ is selected for Gaussian, and $\sigma^2 = 1 \times 10^5$ is selected for M-Gaussian kernel on Yale face database. All these parameters are used in the next section.

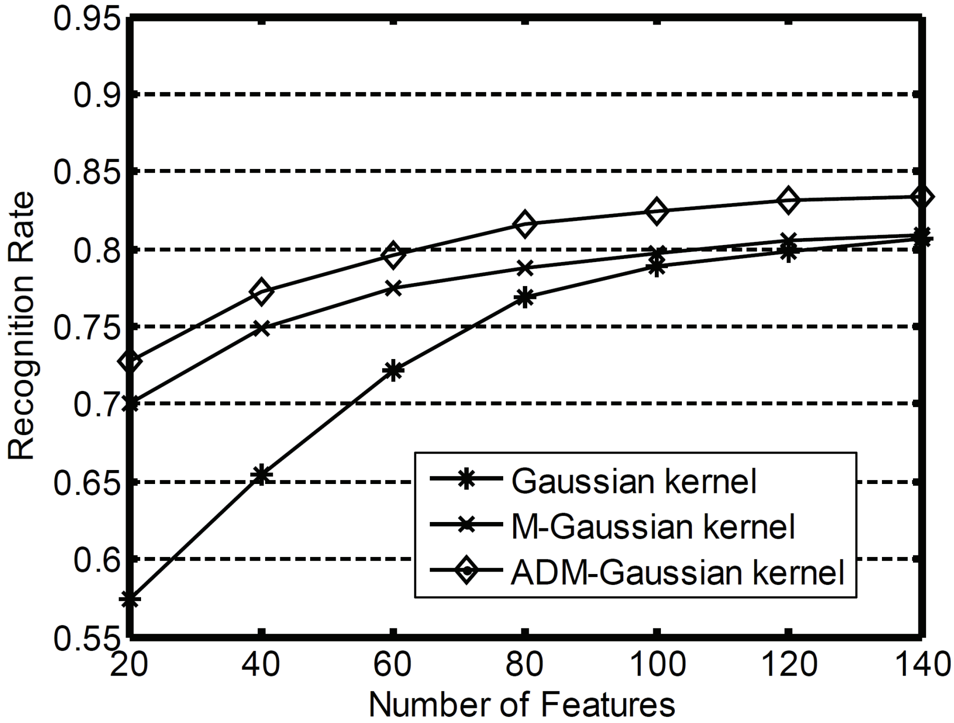


FIGURE 2. Performance on ORL face database

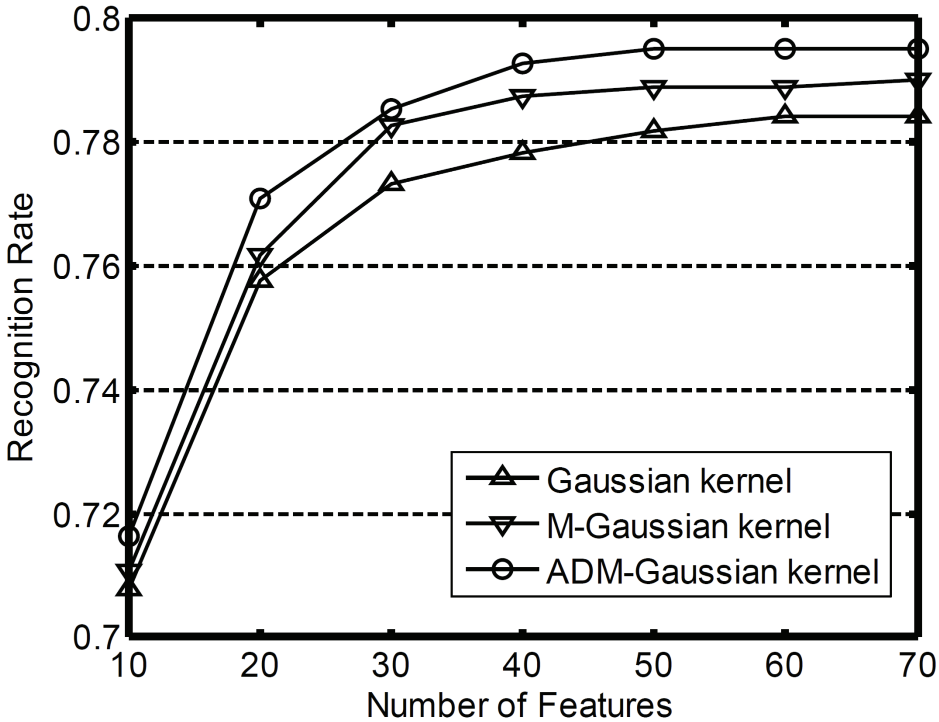


FIGURE 3. Performance on Yale face database

In this section, we evaluate the performance the three kinds of Gaussian kernels based KPCA on the ORL face database and Yale face database. In these experiments, we implement KPCA with the optimal parameters which are selected in the last section. We evaluate the performance of the algorithms with the recognition accuracy with different the dimension of features, and as shown in Fig.2 and Fig.3, we can obtain the highest recognition rate of ADM-Gaussian kernel based KPCA, which is higher than M-Gaussian kernel based KPCA and Gaussian kernel based on KPCA. Moreover the higher recognition rate is obtained with M-Gaussian kernel compared with Gaussian kernel.

M-Gaussian kernel can give the higher recognition accuracy, and owing to the ADM-Gaussian is more adaptive to the input data, which gives a higher recognition accuracy then M-Gaussian. All these experimental results are obtained under the optimal parameters, and the selection of optimal parameters of the original Gaussian kernel will influence the performance the kernel. But the ADM-Gaussian will decrease the influence of the parameter selection by its adaptability for the input data.

4. **Conclusion.** For the feature extraction of image based kernel learning, we present the method of optimizing matrix mapping with data dependent kernel for feature extraction of the image for classification. The method implements the algorithm without transforming the matrix to vector, and it adaptively optimizes the parameter of kernel for nonlinear mapping. The comprehensive experiments are implemented evaluate the performance of the algorithms. A novel kernel namely adaptive data-dependent matrix norm based Gaussian kernel (ADM-Gaussian kernel) is proposed for facial feature extraction. The ADM-Gaussian kernel views images as matrices, which saves the storage and increase the computational efficiency of feature extraction. Adaptive expansion coefficients of ADM-Gaussian kernel are obtained with the maximum margin criterion, which leads to the largest class discrimination of the data in the feature space. The results, evaluated on two popular databases, suggest that the proposed kernel is superior to the current kernel. In the future, we intend to apply the ADM-Gaussian kernel to other areas, such as content-based image indexing and retrieval as well as video and audio classification.

Acknowledgement: This work is supported by the National Science Foundation of China under Grant No. 61001165 and 61371178, the Heilongjiang Provincial Natural Science Foundation of China under Grant No. QC2010066, the HIT Young Scholar Foundation of 985 Project, and the Program for Interdisciplinary Basic Research of Science-Engineering-Medicine at the Harbin Institute of Technology, and the Fundamental Research Funds for the Central Universities (Grant No. HIT.BRETIII.201206).

REFERENCES

[1] Z. M. Lu, X. N. Xu, and J. S. Pan, Face detection based on vector quantization in color images, *International Journal of Innovative Computing, Information and Control*, vol. 2, no. 3, pp. 667-672, 2006.

[2] Y. L. Qiao, Z. M. Lu, J. S. Pan, and S. H Sun, Spline wavelets based texture features for image retrieval, *International Journal of Innovative Computing, Information and Control*, vol. 2, no. 3, pp. 653-658, 2006.

[3] Z. M. Lu, S. Z. Li, and H. Burkhardt, A content-based image retrieval scheme in JPEG compressed domain, *International Journal of Innovative Computing, Information and Control*, vol. 2, no. 4, pp. 831-839, 2006.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[5] A. U. Batur, and M. H. Hayes, Linear subspace for illumination robust face recognition, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 296-301, 2001.

[6] A. M. Martinez, and A. C. Kak, PCA versus LDA, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, 2001.

[7] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in kernel methods: support vector learning*, MIT Press, MA, US, 1999.

[8] A. Ruiz, and P. E. López-de-Teruel, Nonlinear kernel-based statistical pattern analysis, *IEEE Trans. Neural Networks*, vol. 12, no. 1, pp. 16-32, 2001.

[9] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, An introduction to kernel-based learning algorithms, *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181-201, 2001.

[10] B. Schölkopf, A. Smola, and K. R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Journal of Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1998.

[11] Q. Liu, H. Lu, and S. Ma, Improving kernel fisher discriminant analysis for face recognition, *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42-49, 2004.

[12] H. Gupta, A. K. Agrawal, T. Pruthi, C. Shekhar, and R. Chellappa, An experiment evaluation of linear and kernel-based methods for face recognition, *Proc. of the 6th IEEE Workshop on Applications of Computer Vision*, pp.13-18, 2002.

[13] M. H. Yang, Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods, *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 215-220, 2002.

[14] J. W. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Network*, vol. 14, no. 1, pp. 117-126, 2003.

[15] S. Amari, and S. Wu, Improving support vector machine classifiers by modifying kernel functions, *Journal of Neural Network*, vol. 12, no. 6, pp. 783-789, 1999.

[16] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, New York, US, 2000.

[17] H. Li, T. Jiang, and K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 157-165, 2006.

[18] J. B. Li, J. S. Pan, Z. M. Lu, and B. Y. Liao, Data-dependent kernel discriminant analysis for feature extraction and classification, *Proc. of IEEE International Conference on Information Acquisition*, pp.1263-1268, 2006.

[19] F. S. Samaria, A. C. Harter, Parameterisation of a stochastic model for human face identification, *Proc. of the 2nd IEEE Workshop on Applications of Computer Vision*, pp. 138-142, 1994.

[20] G. Malathi, and V. Shanthi, Statistical measurement of ultrasound placenta images complicated by gestational diabetes mellitus using segmentation approach, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 4, pp. 332-343, 2011.

[21] C. F. Lee, and W. T. Chang, Recovery of color images by composed associative mining and edge detection, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 4, pp. 310-324, 2010.

[22] H. G. Kaganami, S. K. Ali, and B. Zou, Optimal approach for texture analysis and classification based on wavelet transform and neural network, *Journal of Information Hiding and Multimedia Signal Processing,* vol. 2, no. 1, pp. 33-40, 2011.

[23] W. C. Hu, C. Y. Yang, D. Y. Huang, and C. H. Huang, Feature-based face detection against skincolor like backgrounds with varying illumination, *Journal of Information Hiding and Multimedia Signal Processing,* vol. 2, no. 2, pp. 123-132, 2011.