Journal of Information Hiding and Multimedia Signal Processing ©2014 ISSN 2073-4212
Ubiquitous International Volume **5**, Number **3**, July **2014**

Kernel-optimized Based Fisher Classification of Hyperspectral Imagery

Xun-Fei Liu, Xiangxian Zhu

Department of Electronic Engineering
Suzhou Institute of Industrial Technology
Suzhou 215104, China
liuxf@siit.edu.cn

Abstract. *This paper is to present a novel kernel-optimized based Fisher classification for hyperspectral imagery. Kernel learning provides a promising solution to the nonlinear problems. The performance of kernel-based learning system has been increased. However, kernel-based system still endures the selection of kernel function and its parameters. Traditional choosing the parameters from a discrete value set did not change the structure of data distribution in kernel-based mapping space. Based on this motivation, we present a uniform framework of kernel self-optimization for kernel-based feature extraction and recognition. In this framework, firstly data-dependent kernel is extended and has a higher ability of adjust the kernel structure, and secondly two criterions are proposed to solve the kernel optimization problem. Evaluations on Two real data sets, namely Indian Pines and Washington, D.C. Mall are implemented to testify the performance of the proposed hyperspectral imagery classification.*

Keywords: Hyperspectral Imagery, kernel learning, Fisher classification

1. **Introduction.** Hyperspectral sensor system has gained popularity in recent decades due to its poten-tial superiority in remote sensing. Among the hyperspectral community, classification of various land covers is one of the fundamental issues, which can particularly benefit from the high spectral resolution of each pixel, compared to the standard images. Among various Hyperspectral imagery classification algorithms, one of the most suc-cessful techniques is the appearance-based method. To resolve the too large dimension problem when using original face images, dimensionality reduction techniques are employed widely [1, 2].Two of the most popular algorithms of these dimensionality reduction techniques are Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2]. Recently, the nonlinear methods, KPCA [7] and KFD [3, 4], have been widely used since kernel machine techniques [5, 6] were applied to the face recognition. The Gabor wavelets, which capture the properties of spatial localization, orientation selectivity, and spatial frequency selectivity to cope with the variations in illumination and facial expressions, are widely employed in face recognition [8, 9]. As the relative works, recently video-based technology have been developed and applied into many research topics including coding [10, 11], enhanc-ing [12, 13] and image processing [14, 15] as discussed in the previous section. In this paper, we propose a novel kernel-optimized based Fisher classification for hyperspectral imagery. The performance of kernel-based learning system has been increased. However, kernel-based system still endures the selection of kernel function and its parameters. Traditional choosing the parameters from a discrete value set did not change the structure of data distribution in kernel-based mapping space. Based

on this motivation, we present a uniform framework of kernel self-optimization for kernel-based feature extraction and recognition. In this framework, firstly data-dependent kernel is extended and has a higher ability of adjust the kernel structure, and secondly two criterions are proposed to solve the kernel optimization problem. Evaluations on Two real data sets, namely Indian Pines and Washington, D.C. Mall are implemented to testify the performance of the proposed hyperspectral imagery classification.

2. **Method.** We apply a novel classification method called Kernel Self-optimization Fisher Dis-criminant (KSFD) for region classification. KSFD method comes from KFD method as follows. The main idea of KFD is to map the original training samples to the feature space F with the nonlinear mapping Φ, and the linear discriminant analysis is implemented in the feature space F. Supposed that $G = diag(G_1, G_2, ..., G_L)$, G_i is a $n_i \times n_i$ matrix consisted of $\frac{1}{n_i}$, K is the kernel matrix calculated by $k(x, y)$. In fact, let $A_{opt} = [\alpha_1, \alpha_2, ..., \alpha_d]$ consisted of d discriminant vector $\alpha_1, \alpha_2, ..., \alpha_d$. The matrix A_{opt} satisfies

$$A_{opt} = \arg\max_{A} \frac{|A^T KGKA|}{|A^T KKA|} \tag{1}$$

Then the feature vector y of sampe x is $y = A_{opt}^T [k(x, x_1), k(x, x_2), ..., k(x, x_n)]^T$.
KFD algorithm:
Step 1. Select kernel function $k\,(x, y)$ and its parameters, calculate the kernel matrix K;
Step 2. Calculate the projection matrix A_{opt};
Step 3. The feature of the sample x is $y = A_{opt}^T [k(x, x_1), k(x, x_2), ..., k(x, x_n)]^T$.
As the above discussion, the projection matrix is the function of kernel matrix. If the kernel function and its parameter are not appropriately chosen, the projection matrix is not optimal for the mapping from the input space to the feature space. So the tradi-tional KFD is not automatically to adjust the data structure in the feature space for feature extraction.

KFD finds an optimal linear projection from the kernel feature space to the projection subspace. Supposed that the nonlinear mapping Φ is inappropriately chosen, KFD can not find the optimal linear projection. In our algorithm, we optimize the nonlinear map Φ to maximize the class separability in feature space by optimizing the kernel, and then find the optimal transformation to maximize the class separability in projection subspace. Based on the above idea, we propose two stages of KFD algorithm, the first one is to optimize the kernel and the second is to find the optimal projection with the traditional method same as KFD. The geometry structure of sample data in the nonlinear projection space is different with the different kernel function. Accordingly, data in the nonlinear projection space has the different class discriminative ability. So the kernel function should be dependent to the input data, which is the main idea of data-dependent kernel. The parameter of the data-dependent kernel is changed according to the input data so that the optimal geometry structure of data in the feature space is achieved for the classification. In this paper, we extend the definition of the data-dependent kernel $k(x, y) = f(x)f(y)k_0(x, y)$ as the objective function for creating the constrained optimization equation to solve the solution, where $k0(x, y)$ is the basic kernel function, such as polynomial kernel and Gaussian kernel. The function $f(x)$ is defined as $f(x) = \sum_{i \in SV} a_i e^{-\delta \|x - \tilde{x}_i\|^2}$, where $\tilde{x}_i$ is the support vector, SV is the set of support vector, a_i denotes the positive value which represent the distribution of $\tilde{x}_i$, δ is the free parameter. We extend the definition of data-dependent kernel through defining the function $f(x)$ with the different ways as $f(x) = b_0 + \sum\limits_{n=1}^{N_{XV}} b_n e(x, \tilde{x}_n)$, where δ is the free parameters, $\tilde{x}_i$ is the expansion vectors (xvs) and N_{XV} is the number of expansion vectors, $bn(n = 0, 1, 2, , NXV)$ is the according expansion coefficients. In our previous work [15], we present four methods of defining $e(x, \tilde{x}_n)$.

Fisher criterion is to measure the class discriminative ability of the samples in the empirical feature space. The discriminative ability of samples in the empirical feature space is defined as

$$J_{Fisher} = \frac{tr(S_B^\Phi)}{tr(S_W^\Phi)} \tag{2}$$

where J_{Fisher} measure the linear discriminative ability, S_B^Φ is the between class scatter matrix, S_W^Φ is inter class scatter matrix, and tr denotes the trace. Let K is the kernel matrix with its element $k_{ij}, (i, j = 1, 2, ..., n)$ is calculated with x_i and x_j. The matrix $K_{pq}, p, q = 1, 2, ..., L$ is the $n_p \times n_q$ matrix with p and q class. Then in the empirical feature space, we can obtain $tr(S_B^\Phi) = 1_n^T B 1_n$ and $tr(S_W^\Phi) = 1_n^T W 1_n$, where $B = diag(\frac{1}{n_1}K_{11}, \frac{1}{n_2}K_{22}, ..., \frac{1}{n_L}K_{LL}) - \frac{1}{n}K$. The class discriminative ability is defined as

$$J_{Fisher} = \frac{1_n^T B 1_n}{1_n^T W 1_n} \tag{3}$$

According to the definition of the data-dependent kernel, let $D = diag(f(x_1), f(x_2), ..., f(x_n))$, the relation between the data-dependent kernel matrix K and the basic kernel matrix K_0 calculated with basic kernel function $k_0(x, y)$ is defined as $K = DK_0D$. Accordingly, $B = DB_0D$ and $W = DW_0D$. Then

$$J_{Fisher} = \frac{1_n^T DB_0D 1_n}{1_n^T DW_0D 1_n} \tag{4}$$

where 1_n is n dimensional unit vector, according to the definition of data-dependent kernel, then

$$D1_n = E\alpha \tag{5}$$

where $\alpha = [a_0, a_1, a_2, ..., a_{N_{XVs}}]^T$, the matrix E is consisted of $e(x, \tilde{x}_n)$. Then

$$J_{Fisher} = \frac{\alpha^T E^T B_0 E\alpha}{\alpha^T E^T W_0 E\alpha} \tag{6}$$

where $E^T B_0 E$ and $E^T W_0 E$ are constant matrix, J_{Fisher} is a function with its variable α. Under the different expansion coefficient vector α, the geometry structure of data in the empirical space causes the discriminative ability of samples. Our goal is to find the optimal α to maximize J_{Fisher}. Supposed that α is an unit vector, i.e., $\alpha^T\alpha = 1$, the constrained equation is created to solve the optimal α as follows.

$$\begin{array}{rl} \max & J_{Fisher}(\alpha) \\ subject \quad to & \alpha^T\alpha - 1 = 0 \end{array} \tag{7}$$

There are many methods of solving the above optimization equation. The following method is a classic method. Let $J_1(\alpha) = \alpha^T E^T B_0 E\alpha$ and $J_2(\alpha) = \alpha^T E^T W_0 E\alpha$, then

$$\begin{cases} \frac{\partial J_1(\alpha)}{\alpha} = 2E^T B_0 E\alpha \\ \frac{\partial J_2(\alpha)}{\alpha} = 2E^T W_0 E\alpha \end{cases} \tag{8}$$

Then

$$\frac{\partial J_{Fisher}(\alpha)}{\partial\alpha} = \frac{2}{J_2^2}(J_2 E^T B_0 E - J_1 E^T W_0 E)\alpha \tag{9}$$

In order to maximize J_{Fisher}, let $\frac{\partial J_{Fisher}(\alpha)}{\partial\alpha} = 0$, then

$$J_1 E^T W_0 E\alpha = J_2 E^T B_0 E\alpha \tag{10}$$

If $\left(E^T W_0 E\right)^{-1}$ exists, then

$$J_{Fisher}\alpha = (E^T W_0 E)^{-1}(E^T B_0 E)\alpha \tag{11}$$

J_{Fisher} is equal to the eigenvalue of $(E^T W_0 E)^{-1}(E^T B_0 E)$, and the corresponding eigenvector is equal to expansion coefficients vector α. In many applications, the matrix $(E^T W_0 E)^{-1}(E^T B_0 E)$ is not symmetrical or $E^T W E$ is singular. So the iteration method is to solve α as follows.

$$\alpha^{(n+1)} = \alpha^{(n)} + \varepsilon(\frac{1}{J_2}E^T B_0 E - \frac{J_{Fisher}}{J_2}E^T W_0 E)\alpha^{(n)} \tag{12}$$

ε is the learning rate as follows. The definition of learning rate is

$$\varepsilon\left(n\right) = \varepsilon_0(1 - \frac{n}{N}) \tag{13}$$

where ε_0 is the initialized learning rate, n and N is the current iteration number and the total iteration number in advance respectively.

The initialized learning rate ε_0 and the total iteration number N is set in advance for the solution of the expansion coefficient. The initial learning rate ε_0 influences the convergence speed of the algorithm, and the total iteration number N determines the time of solution. Only when the parameter ε_0 and N are chosen appropriately we choose the optimal expansion coefficient vector. So the solution of expansion coefficient is not unique, which is determined by the selection of learning parameter. The iteration algorithm costs much time. So we select the maximum margin criterion as the objective function to solve the optimal expansion coefficients.

3. Experimental results.

3.1. Results on simulated data.

We test the proposed kernel optimization methods through simulation, which are implemented on simulated dataset and two face databases. We use the basic kernel function for the nonlinear mapping from the input space into the empirical feature space. Figure 1 shows the data distribution in the empirical feature space with the basic kernels of Gaussian kernel and Polynomial kernel. The discriminative ability of samples decreased if the kernel function is not good as shown in Figure 2, so kernel optimization is necessary. Figure 3 shows the comparison of Fisher criterion and maximum margin criterion, where the iteration number of Fisher method is set to 400. The data in the empirical feature space is similar with Fisher method and maximum margin criterion method, but the Fisher criterion method is influenced by parameter setting, such as the learning rate and iteration number. So the maximum margin criterion out-performs Fisher criterion method. As the above discussion, kernel function influences the performance of kernel learning. If the kernel function and its parameters are inappropriately chosen the performance will decrease. Two kernel optimization methods, Fisher criterion and maximum margin criterion methods, achieve the similar performance of kernel optimization, but Fisher method endures the solution problem.

3.2. Results on two real datasets.

Two real data sets, namely Indian Pines and Washington, D.C. Mall, with various spectral and spatial resolutions reecting dfierent environments of remote sensing are adopted in the experiments.Two real data sets, namely Indian Pines and Washington, D.C. Mall, with various spectral and spatial resolutions reflecting different environ-ments of remote sensing are adopted in the experiments.

1) Indian Pines data: the first test set to be used was the well-known Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) image scene, which was captured over the agricultural region of Northwestern Indiana in June 1992, with spectral resolution of

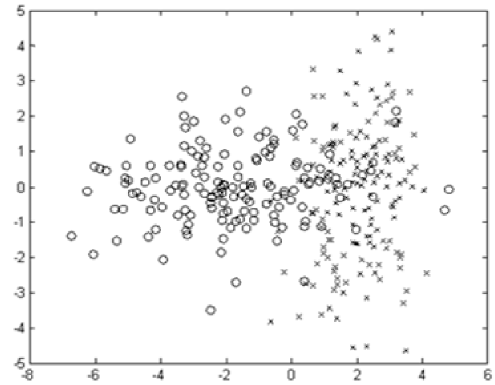
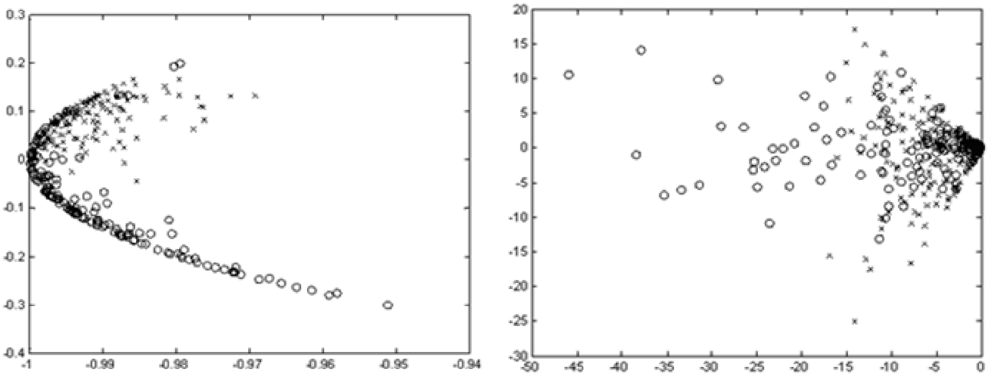


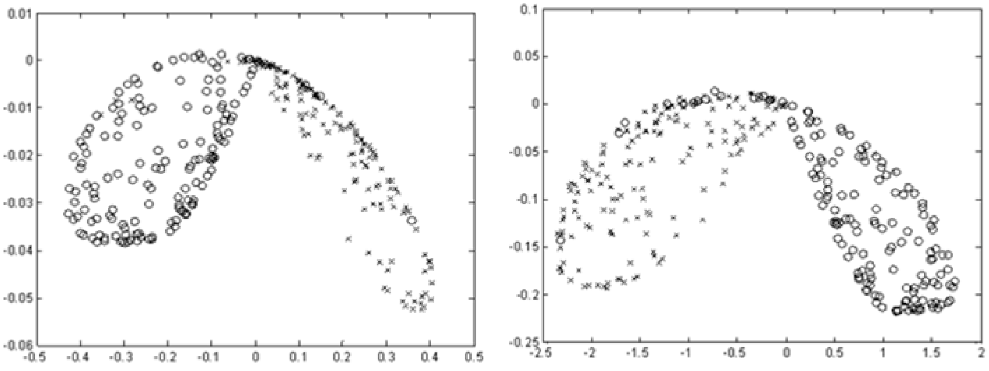
Figure 1. Two classes of two-dimensional data samples with Gaussian distribution



(a) Gaussian kernel

(b) Polynomial kernel

Figure 2. Distribution of data samples in the empirical feature space



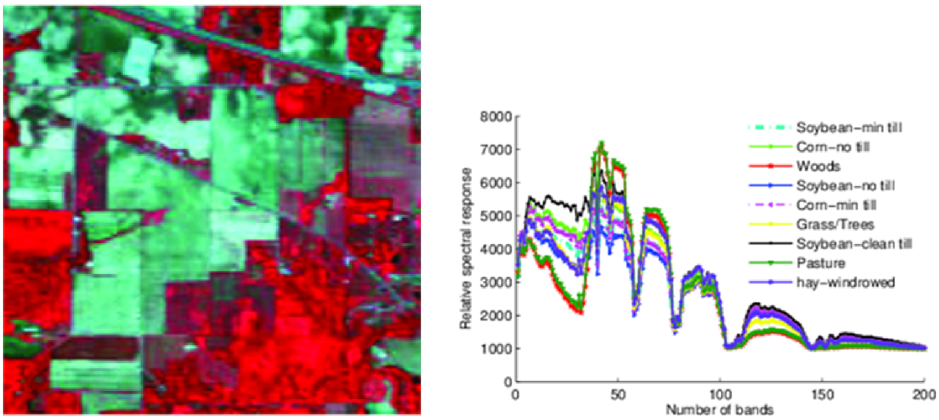
(a) Maximum margin criterion

(b) Fisher criterion

Figure 3. Performance comparison of two algorithms

224 bands covering the 0.4-2.5 μm range and spatial resolution of 20m per pixel. After removing the noisy and water-vapor absorption bands, 200 bands reserved for experiments. Although the whole scene consists of 145×145 pixels with 16 classes of interest, ranging the size from 20 to 2468 pixels, only 9 classes with high number of samples are selected. Additionally, the false color image and spectral signatures are depicted in Figure 4 .

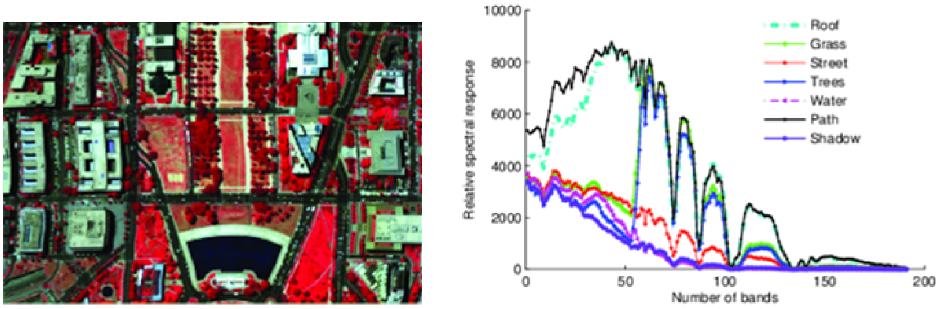
2) D.C. Mall data: the second test set was acquired by the airborne hyperspectral digital imagery collection experiment (HYDICE) sensor over a Mall in Washington D.C.



(a) Three band false color composite (b) Spectral signatures

FIGURE 4. Indian Pines data

on August 23, 1995. The whole urban image size is 1280×307 pixels with the spatial resolution of 1.5m by pixel and 210 spectral bands in the 0.4-2.4 μm region. Several undesirable bands influenced by the atmospheric absorption are discarded, leaving 191 bands for experiments. From the original image, we crop a [870-1080]×[1-307] subset with a size of 211×307, which composed of 7 classes of land-covers (i.e. roof (o_1), grass (o_2), street (o_3), trees (o_4), water (o_5), path (o_6) and sha-dow (o_7)). Table 7 exhibits the detailed number of each class utilized in the experi-ments. The false color image together with spectral signatures are displayed in Figure 5.



(a) Three band false color composite (b) Spectral signatures

FIGURE 5. D.C. Mall data

It is notable that all the experimental results are the mean accuracies over 10 repetitions, which helps to compare different methods in a fair and reasonable way. Without loss of generality, the classification maps of a trial with both HSI data sets are de-picted, where the classifiers with the kernels are illustrated on the bottom of each map. One can roughly observe that with the same classifiers, the KSFC based methods ex-hibit lower classification errors than other ones for both data sets. Observed from Tables 1-2, four main results can be highlighted: 1) Among all algorithms, the OMP-based SRC without any kernels lead to the worst classification accuracies for both data sets. 2) With the same classifiers (i.e. SVM or SRC), results of the KSFC are the best, the RBF is next, whereas those of the POL are the lowest. 3) With the same ker-nels (i.e. POL, RBF or KSFC), the SRC reveals slightly better performance than the SVM. 4) With the same data sets (i.e. Indian Pines data or D.C. Mall data), the com-putation time of the KSFC based classifiers is longer than other ones, which is accept-able since the MKL is quite time consuming in dealing with HSI. Moreover, the SVM is much fast than the SRC in case the same kernels are provided. In a nutshell, the KSFC achieves better classification results than the widespread POL and RBF within moderate time.

Table 1. Two classes of two-dimensional data samples with Gaussian distribution

Class	SVM(POL)	SVM(RBF)	SVM(KSFC)	SRC(OMP)	SRC(POL)	SRC(RBF)	KSFC
1	48.04	77.19	77.82	46.17	50.85	77.70	80.45
2	59.34	71.38	79.78	58.91	59.36	76.45	84.24
3	96.60	98.88	99.91	94.12	96.26	99.12	99.45
4	24.48	72.97	80.93	34.55	47.56	75.66	83.56
5	26.67	75.84	90.02	72.13	75.32	78.73	93.67
6	93.25	97.11	99.24	95.43	93.76	97.44	99.64
7	63.11	77.86	82.54	69.88	62.78	82.76	84.44
8	85.03	89.01	98.81	90.31	85.16	88.79	97.92
9	100	99.10	100	99.77	100	100	100

Table 2. Two classes of two-dimensional data samples with Gaussian distribution

Class	SVM(POL)	SVM(RBF)	SVM(KSFC)	SRC(OMP)	SRC(POL)	SRC(RBF)	Proposed
o1	77.79	86.82	92.57	77.69	79.83	90.63	96.23
o2	96.64	96.81	98.21	96.85	97.18	98:21	97.24
o3	93.01	95.72	97.73	90.19	91.37	97.32	99.22
o4	95.53	96.45	97:89	95.24	96.16	97.79	97.24
o5	100	100	99.86	100	99.86	100	99.55
o6	100	100	99.23	100	100	89.97	94.64
o7	94.19	95.67	97.54	94.12	93.21	97.54	97.30

4. **Conclusions.** In this paper, we present a novel kernel-optimized based Fisher classification for hyperspectral imagery. The proposed classifier is applied to classification. The expe-rimental results on two real data sets, namely Indian Pines and Washington, D.C. Mall show that the proposed algorithm is effective.

Acknowledgment. This work is supported by Basic Construction Project of Science and Technology in Suzhou (Grand No. SZP201212), and Small and medium-sized enterprises in Suzhou city public service platform of Embedded Technology.

REFERENCES

[1] 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.

[2] 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. II-296-II301, 2001.

[3] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230-244, 2005.

[4] Q. S. Liu, H. Q. Lu, and S. D. Ma, Improving kernel fisher discriminant analysis for face recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 1, pp. 42-49, 2004.

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

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

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

[8] C. Liu, Gabor-based kernel PCA with fractional power polynomial models for face recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572-581, 2004.

[9] C. Liu, and H. Wechsler, Independent component analysis of gabor features for face recognition, *IEEE Trans. Neural Networks*, vol. 14, no. 4, pp. 919-928, 2003.

[10] H. Wang, J. Liang, and C. C. Jay Kuo. Overview of robust video streaming with network coding, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 1, pp. 36-50, 2010.

[11] J. Lou, S. Liu, A. Vetro, and M. T. Sun, Trick-play optimization for H.264 video decoding, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 2, pp. 132-144, 2010.

[12] Y. B. Rao, and L. T. Chen, A survey of video enhancement techniques, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 3, no. 1, pp. 71-99, 2012.

[13] Y. B. Rao, and L. T. Chen, An efficient contourlet-transform-based algorithm for video enhancement, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 3, pp. 282-293, 2011.

[14] Á. Serrano, I. Martín de Diego, C. Conde, and E. Cabello, Recent advances in face biometrics with gabor wavelets: a review, *Pattern Recognition Letters*, vol. 31, no. 5, pp. 372-381, 2010.

[15] M. Parviz, and M. S. Moin, Boosting approach for score level fusion in multimodal biometrics based on AUC maximization, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 1, pp. 51-59, 2011.