Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **5**, September **2015**

Parallel Feature Extraction through Preserving Global and Discriminative Property for Kernel-Based Image Classification

Xun-Fei Liu, and Xiang-Xian Zhu

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

Received May, 2015; revised June, 2015

Abstract. *Kernel-based feature extraction is widely used in image classification, and different kernel methods extract the features based different criterion. KPCA maximizes the determinant of the total scatter matrix of the transformed sample, while KDA seeks the direction of discrimination. KPCA preserves the global property, and KDA utilizes class information to enhance its discriminative ability so as to perform better than KPCA in classifications. To apply the global property and discriminant ability of features, we propose a novel parallel feature fusion method based maximum margin criterion, namely discriminant parallel feature fusion. The advantage of algorithm lies in: 1) A constrained optimization problem based on maximum margin criterion is created to solve the optimal fusion coefficients to be most discriminant in the fused feature space. 2) An unique solution of optimization problem is transformed to an eigenvalue problem, which causes the proposed fusion strategy to perform consistently. Besides of the detailed theory derivation, many experimental evaluations also are presented in this paper.*
Keywords: Kernel learning, Image classification, Discriminant parallel feature fusion

1. **Introduction.** Dimensionality reduction (DR) is the most popular approach for feature extraction. DR has wide applications in computer vision, pattern recognition, gene expression, paleontology, etc. 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], enhancing [12, 13] and image processing [14, 15] as discussed in the previous section.

How to perform a wonderful classification based on the multiple features becomes a crucial problem for pattern classification problem when multiple features are considered. As a very efficient method, data fusion is applied to solve it, which has been widely applied in many areas [16, 17, 18]. Existing fusion methods can be divided into the following three schemes: the first scheme is to integrate all assimilated multiple features

into a final decision directly; the second is to combine the individual decisions made by every feature into a global final decision; and the third is to fuse the multiple features to one new feature for classification. In this paper, we devote our attention to the third fusion scheme, i.e., feature fusion. Recently many feature fusion methods for pattern classification were proposed in the lectures [19, 20, 21].

In this paper, we focus on the linear combination fusion, but pay more attention to how to find the fusion coefficients, and propose so called discriminant feature fusion for supervising learning. The proposed discriminant fusion strategy has two advantages: 1) fused data has the largest class discriminant owing to obtaining the fusion coefficients by solving a constrained optimization problem created in the average margin criterion; 2) fusion coefficients are unique owing to they are equal to the elements of the eigenvector of one eigenvalue problem transformed by the above optimization problem. Moreover, multiple kernel based combination learning methods are developed including Sparse Multiple Kernel Learning [25], Large Scale Multiple Kernel Learning [26], Lp-Norm Multiple Kernel Learning [27], and on hyperspectral image classification [28]. These methods were reported an excellent performance on feature extraction, classification of data analysis. These methods applied the unchangeable combination parameters during kernel learning machine. So the combined kernel structure is not changed with the distribution structure of the data.

2. **Kernel-Mapping Dimensionality Reduction.**

In this section we review and analyze the kernel discriminant analysis (KDA), locality preserving projection (LPP) and kernel principal component analysis (KPCA).

2.1. **KDA.**

KDA transforms the transformation matrix from the input space to a nonlinear high-dimensional feature space [22]. Given L classes of M training samples $\{x_1, x_2, ..., x_M\}$ in an N-dimensional space R^N, the data are mapped into a feature space F via the following nonlinear mapping:

$$\Phi : R^N \to F, xa\ \Phi(x) \tag{1}$$

The Fisher criterion in the feature space F is defined by

$$J(V) = \frac{V^T S_B^\Phi V}{V^T S_T^\Phi V} \tag{2}$$

where V is the discriminant vector, and S_B^Φ and S_T^Φ are the between-class scatter matrix and total-scatter matrix, respectively. Any solution V belongs to the span of all training patterns in R^N. Hence, there exists coefficients $c_p(p = 1, 2, ..., M)$ such that

$$V = \sum_{p=1}^{M} c_p \Phi(x_p) = \Psi\alpha \tag{3}$$

where $\Phi = [\Phi(x_1), \Phi(x_2), ..., \Phi(x_M)]$ and $\alpha = [c_1, c_2, ..., c_M]^T$. Assuming that the data are centered, the Fisher criterion is transformed into

$$J(\alpha) = \frac{\alpha^T KGK\alpha}{\alpha^T KK\alpha} \tag{4}$$

where $G = diag(G_1, G_2, ..., G_L)$, G_i is an $n_i \times n_i$ matrix whose elements are $\frac{1}{n_i}$, and the kernel matrix K is calculated by a basic kernel $k(x, y)$.

2.2. **LPP.** As a dimensionality reduction method, the locality preserving projection (LPP) preserves the locality of the data during reducing the dimensionality. Given m classes of N-dimensional data $\{x_1, x_2, ..., x_n\}$, the LPP aims to find a transformation matrix W to map the N-dimensional sample vector to a lower-dimensional dataset $\{z_1, z_2, ..., z_n\}$. The objective function of LPP is defined as follows [22]:

$$min \sum_{i,j} \|w^T x_i - w^T x_j\|^2 S_{ij} \ subject\ to\ w^T w = 1 \tag{5}$$

where S is a similarity matrix whose elements measure the similarity of two points. By minimizing the objective function in (5), LPP incurs a heavy penalty if the neighboring mapped points, z_i and z_j, are far. However, it keeps the mapped points close if the original points are close. Minimizing (5) is equivalent to the following equation:

$$\frac{1}{2} \sum_{i,j} \|z_i - z_j\|^2 S_{ij} = w^T X(D - S)X^T w = w^T XLX^T w \tag{6}$$

where $X = [x_1, x_2, ..., x_n]$, and D is a diagonal matrix whose entries are the column or row (S is symmetric) sums of S; i.e., ,$D = diag\left[\sum_j S_{1j}, \sum_j S_{2j}, ..., \sum_j S_{nj}\right]$ and $L = D - S$ is the Laplacian matrix. Matrix D describes the local structure information of the data. A constraint is imposed as follows:

$$z^T Dz = 1 \Rightarrow w^T XDX^T w = 1 \tag{7}$$

Then,

$$\underset{w}{arg\ min}\ \ w^T XLX^T w \ s.t.\ w^T XDX^T w = 1 \tag{8}$$

The optimal transformation vector w is computed through the solving of the eigenvalue problem,

$$XLX^T w = \lambda XDX^T w \tag{9}$$

where $L = D - S$, and $D = diag\left[\sum_j S_{1j}, \sum_j S_{2j}, ..., \sum_j S_{nj}\right]$. The similarity matrix S is defined as

$$S_{ij} = \begin{cases} exp\Big(-\frac{1}{k}\|x_i - x_j\|^2\Big) & if\ x_i\ is\ one\ of\ k\ nereast\ neighbors\ of\ x_j \\ & or\ x_j\ is\ one\ of\ k\ nereast\ neighbors\ of\ x_i \\ 0 & \end{cases} \tag{10}$$

otherwise, the rank of XDX^T is at most n , while XDX^T is an $N \times N$ matrix. The matrix XDX^T is singular. LPP employs a procedure that is similar to Fisherface [17] to overcome the singularity of XDX^T . The procedure of LPP is described as follows: Step 1: Project data with the projection matrix W_{PCA} . Step 2: Construct the nearest-neighbor graph G and the similarity matrix S . Step 3: Calculate $W = W_{PCA}W_{LPP}$, where W_{LPP} is the LPP projection matrix.

2.3. **KPCA.** KPCA is a popular dimensionality reduction method as a kerenelized version of PCA. For a clear description, we introduce PCA, which is then kernelized into KPCA [15]. Given the training samples $x_1, x_2, ..., x_n$, $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$, where *overlinex* is the mean of all of the training samples, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Generally, the SVD is applied to solving the singular matrix problem. If $Q = [x_1 - \overline{x}, ..., x_n - \overline{x}]$, then $C = \frac{1}{n}QQ^T$, and $R = Q^T Q$ is the $n \times n$ positive definite matrix. The dimension of R is less than

that of C . For the eigenvectors $V_1, V_2, ..., V_m$ according to the m largest values ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$), the projection vectors are computed via $w_j = \frac{1}{\sqrt{\lambda_j}} Q v_j, j = 1, 2, ..., m$. For any sample x , the j th feature is $y_j = w_j^T x = \frac{1}{\sqrt{\lambda_j}} v_j^T Q^T x, j = 1, 2, ..., m$. PCA is kernelized as follows. $c = \frac{1}{n} \sum_{i=1}^{n} (\Phi(x_i) - \overline{\Phi})(\Phi(x_i) - \overline{\Phi})^T$, where $\overline{\Phi} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)$, and if $C' = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T$ and $Q = [\Phi(x_1), ..., \Phi(x_n)]$, then $C' = frac1nQQ^T$ according to $R' = Q^T Q$ using the kernel function. If the eigenvectors $u_1, u_2, ..., u_m$ are computed according to the m th eigenvalue $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ of R , then $w_1, w_2, ..., w_m$ is calculated by $w_j = \frac{1}{\sqrt{\lambda_j}} Q u_j, j = 1, 2, ..., m$ and $y_j = w_j^T x = \frac{1}{\sqrt{\lambda_j}} u_j^T [k(x_1, x), k(x_2, x), ..., k(x_n, x)]$

2.4. **Discussion.** KPCA maximizes the determinant of the total scatter matrix of the transformed sample, while KDA seeks the direction of discrimination. Both KPCA and LPP are unsupervised learning methods, while KDA is a supervised learning method. KPCA preserves the global property, while LPP preserves the local structure. As a global method, KDA utilizes class information to enhance its discriminative ability, and therefore, it performs better than KPCA in assigning classifications.

3. **Discriminant parallel feature fusion.** In this section, firstly we introduce the basic idea of discriminant parallel feature fusion briefly, and then emphasize the theory derivation of seeking the fusion coefficients in detailed. On current machine learning methods, the performance of many linear learning methods is improved because the data distribution in the nonlinear feature space is easy to classification owing to data mapping. The geometrical structure of the data in the mapping space, which is totally determined by the mapping model, has significant impact on the performance of these learning methods. The discriminative ability of the data in the feature space could be even worse if an inappropriate model is used. So, we present a novel discriminant kernel fusion method, and in this framework a constrained optimization problem based on maximum margin criterion is created to solve the optimal fusion coefficients, which causes that fused data has the largest class discriminant in the fused feature space. An unique solution of optimization problem is transformed to an eigenvalue problem, which causes the proposed fusion strategy to perform a consistent performance.

Given a sample set $x_{ij} (i = 1, 2, ..., C; j = 1, 2, ..., n_i)$, and multiple feature sets $y_{ij}^m (i = 1, 2, ..., C; j = 1, 2, ..., n_i; m = 1, 2, ..., M)$, where M denotes the number of multiple features sets, the fused feature with the linear combination can be described as follows.

$$z_i^j = \sum_{m=1}^{M} a_m y_{ij}^m \tag{11}$$

where $a_m (m = 1, 2, ..., M)$ and $z_{ij} (i = 1, 2, ..., C; j = 1, 2, ..., n)$ denote the combination fusion coefficients and the fused feature respectively. Now we focus how to obtain $a_m (m = 1, 2, ..., M)$, and our goal is to find such fusion coefficients that they are unique and cause the largest class discriminant in the fused feature space. For supervised learning, we can calculate the average margin distance between two classes C_p^1 and C_q^1 in fused feature space consisted of the fused feature$z_i^j = y_i^j \alpha (i = 1, 2, \ldots, C; j = 1, 2, \ldots, n_i)$, where $\alpha = [a_1, a_2, \ldots, a_M]^T$ and $y_i^j = [y_{ij}^1, y_{ij}^2, \ldots, y_{ij}^M]$. The average margin distance can be

defined by

$$Dis = \frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q d\left(C_p^1, C_q^1\right) \tag{12}$$

where $d\left(C_p^1, C_q^1\right)$ denotes the margin distance between pth and qth classes. Given the feature vector z_i^j in the dimension-reduced space F^1 , and $m_i^1 (i = 1, 2, \ldots, L)$ and m_i^1 denote the mean of every class and the mean of total samples respectively. Firstly we can calculate $d\left(C_p^1, C_q^1\right) (p = 1, 2, ..., L; q = 1, 2, \ldots, L)$ as follows.

$$d\left(C_p^1, C_q^1\right) = d\left(m_p^1, m_q^1\right) - S(C_p^1) - S(C_q^1) \tag{13}$$

where $S(C_p^1)(p = 1, 2, \ldots, L)$ is the measure of the scatter of the class C_p^1 and $d\left(m_p^1, m_q^1\right)$ is the distance between the means of two classes. Let $S_p^1(p = 1, 2, \ldots, L)$ denote the within-class scatter matrix of class p, then $tr(S_p^1)(p = 1, 2, \ldots, L)$ measures the scatter of the class p can be defined as follows.

$$tr(S_p^1) = \frac{1}{n_p}\sum_{j=1}^{n_p}(z_p^j - m_p^1)^T(z_p^j - m_p^1) \tag{14}$$

And we can define $tr(S_B^1)$ and $tr(S_W^1)$ denote the trace of between classes scatter matrix and within classes scatter matrix of dimension-reduced space F^1respectively as follows.

$$tr(S_B^1) = \sum_{p=1}^{L} n_p(m_p^1 - m^1)^T(m_p^1 - m^1) \tag{15}$$

$$tr(S_W^1) = \sum_{p=1}^{L}\sum_{j=1}^{n_i}(z_p^j - m_p^1)^T(z_p^j - m_p^1) \tag{16}$$

Hence $S(C_p^1) = tr(S_p^1)$. So

$$Dis = \frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q \left[d(m_p^1, m_q^1) - S(C_p^1) - S(C_q^1)\right]$$

$$= \frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q d(m_p^1, m_q^1) - \frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q \left[tr(S_p^1) + tr(S_q^1)\right] \tag{17}$$

Firstly we use Euclidean distance to calculate $d(m_p^1, m_q^1)$ as follows.

$$\frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q d(m_p^1, m_q^1) = \frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q (m_p^1 - m_q^1)^T(m_p^1 - m_q^1) \tag{18}$$

According to equation (5), (6), (8) and (9), it is easy to obtain

$$\frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q d(m_p^1, m_q^1) = tr(S_B^1) \tag{19}$$

$$\frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q \left[tr(S_p^1)\right] = \frac{1}{2}tr(S_W^1) \tag{20}$$

Hence

$$\frac{1}{2n}\sum_{p=1}^{L}\sum_{q=1}^{L} n_p n_q \left[tr(S_p^1) + tr(S_q^1)\right] = tr(S_W^1) \tag{21}$$

. We can obtain

$$Dis = tr(S_B^1) - tr(S_W^1) \tag{22}$$

In the previous work in [22], Li applied the maximum margin criterion to feature extraction by maximizing the average margin distance. In this paper, we expect to create an optimization problem based on maximum margin criterion to seek an optimal projection vector α.

Proposition 1. Let

$$G = 2\sum_{i=1}^{L}\frac{1}{n_i}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}(y_i^j)^T y_i^k\right] - \sum_{i=1}^{L}\sum_{p=1}^{L}\sum_{j=1}^{n_i}\sum_{qj=1}^{n_p}\left(\frac{1}{n}(z_i^j)^T y_p^q\right) - \sum_{i=1}^{L}\sum_{j=1}^{n_i}(y_i^j)^T y_i^j \tag{23}$$

Then $Dis = \alpha^T G\alpha$.

Proof. From equation (5) and (6), we can obtain

$$tr(S_B) = \sum_{i=1}^{L}\frac{1}{n_i}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}(y_i^j)^T y_i^k\right] - \sum_{i=1}^{L}\sum_{p=1}^{L}\sum_{j=1}^{n_i}\sum_{qj=1}^{n_p}\left(\frac{1}{n}(z_i^j)^T z_p^q\right) \tag{24}$$

$$tr(S_W) = \sum_{i=1}^{L}\sum_{j=1}^{n_i}(z_i^j)^T z_i^j - \sum_{i=1}^{L}\frac{1}{n_i}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left(z_i^j\right)^T z_i^k\right] \tag{25}$$

From equation (2) $z_i^j = y_i^j\alpha$, we can obtain Let

$$G = 2\sum_{i=1}^{L}\frac{1}{n_i}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}(y_i^j)^T y_i^k\right] - \sum_{i=1}^{L}\sum_{p=1}^{L}\sum_{j=1}^{n_i}\sum_{qj=1}^{n_p}\left(\frac{1}{n}(z_i^j)^T y_p^q\right) - \sum_{i=1}^{L}\sum_{j=1}^{n_i}(y_i^j)^T y_i^j \tag{26}$$

, It is easy to obtain

$$tr(S_B^\Phi) - tr(S_W^\Phi) = \alpha^T G\alpha \tag{27}$$

According to Propositions 1, we can obtain $Dis = \alpha^T G\alpha w$

According to the maximum margin criterion and Proposition 1, we can create an optimization problem constrained by the unit vector α, i.e., $\alpha^T\alpha = 1$, as follows.

$$\max_{\alpha}\ \alpha^T G\alpha \tag{28}$$

subject to

$$\alpha^T\alpha - 1 = 0 \tag{29}$$

In order to solve the above constrained optimization equation, we apply a Lagrangian

$$L(\alpha,\lambda) = \alpha^T G\alpha - \lambda(\alpha^T\alpha - 1) \tag{30}$$

with the multiplierλ. The derivative of $L(\alpha,\lambda)$ with respect to the primal variables must vanish, that is

$$\frac{\partial L(\alpha,\lambda)}{\partial\alpha} = (G - \lambda I)\alpha = 0 \tag{31}$$

$$\frac{\partial L(\alpha,\lambda)}{\partial\lambda} = 1 - \alpha^T\alpha = 0 \tag{32}$$

Hence

$$G\alpha = \lambda\alpha \tag{33}$$

The detail optimal parameter vectors is defined to denote the class discriminative ability of the data, and the ability is measured by maximum margin criterion, which is created

to solve the optimal fusion coefficients, which causes that fused data has the largest class discriminant in the fused feature space. An unique solution of optimization problem is transformed to an eigenvalue problem, which causes the proposed fusion strategy to perform a consistent performance.

The problem of solving the constrained optimization function is transformed to the problem of solving eigenvalue equation shown in (19). The fusion coefficients are equal to the elements of eigenvector of corresponding to the largest eigenvalue, while is a matrix which can be calculated by multiple features.

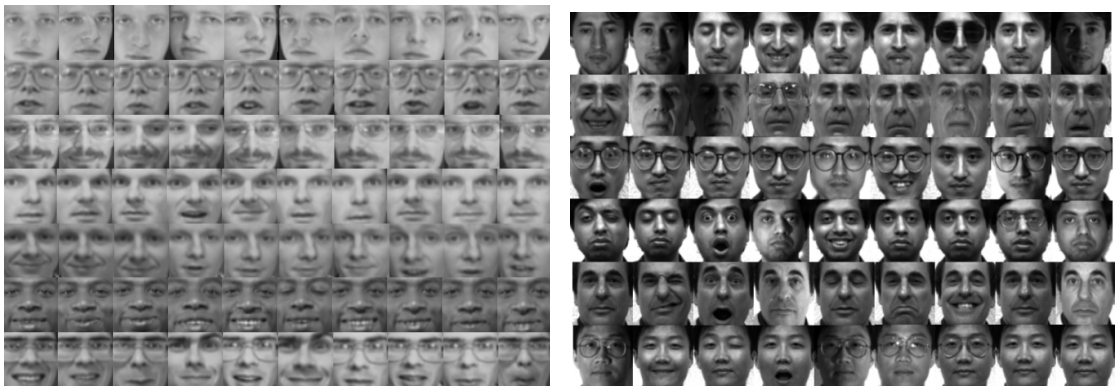
As above discussion, discriminant feature fusion finds a discriminating fused feature space, in which data has largest class discriminant. And then the fusion coefficients are equal to the elements of eigenvector of an eigenvalue problem corresponding to the largest eigenvalue, and the solution of the eigenvalue problem is unique, so the fusion coefficients are unique.

4. **Experimental results.** In our experiments, we implement our algorithm in the two face databases, ORL face database [24] and Yale face database [23]. 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.

The experimental procedural parameter In the practical applications are chosen as follows. The kernel function is adaptively chosen subject to the training samples set. The training sample set is constructed by the training images, and the kernel function can be adaptively chosen by kernel machine. In the experiments, we choose the procedural parameters through cross-validation method for the practical application. All training samples are considered the samples to cross-validation method. In the practical applications, we choose it with expert experience for some applications. And the detailed parameter will be solved through optimizing the constrained equations.

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 of the algorithms.

From the theory derivation of discriminant fusion in Section 2, it is easy to predict that the proposed algorithm gives the better performance compared with the classical fusion [4], and here only a set of experiments are implemented for evaluation. Firstly we extract the linear and nonlinear features with KDA and KPCA, and then classify the fused feature of the two features with Fisher classifier. Supposed y_{ij}^1 and $y_{ij}^2 (i = 1, 2, \ldots, C;\ j = 1, 2, \ldots, n_i)$ denote the linear and nonlinear feature derived from PCA and KPCA respectively, the fused feature, $z_i^j = \begin{pmatrix} y_{ij}^1 \\ y_{ij}^2 \end{pmatrix}$ based on the classical fusion [4], while $z_i^j = \sum_{m=1}^{2} a_m y_{ij}^m$ based on discriminant fusion strategy. Here Polynomial kernel and



(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).

Gaussian kernel with different coefficients are selected for KPCA, and accuracy rate is applied to evaluate the recognition performance.

As Table 1, 2, 3, 4 shown, the proposed method gives a higher performance than the classical fusion method [4] under the same kernel parameters for KPCA.

Since the fusion coefficients of the discriminant fusion strategy are obtained by solving the optimization problem based on maximum margin criterion and data has the largest class discriminant in the fused feature space, it is not surprised that discriminant fusion gives a consistently better performance than classical fusion [4]. But besides the above advantages, the following cases are worthy to be considered: 1) Since the maximum margin criterion is used to create the constrained optimization problem, the fusion strategy is only adapted to the supervised learning. 2) The fusion coefficients are obtained by solving one eigenvalue problem, which causes the increasing of time consuming than classical strategy, so it should evaluate the balance of time consuming and classification accuracy. 3) Discriminant fusion strategy is only a linear combination of multiple features with different combination coefficients, so other fusion strategies can be considered to create based on the discriminant analysis.

TABLE 1. Performance on ORL face database. (Polynomial kernel for KPCA)

Methods	d=2	d=3	d=4	d=5
Fusion method[4]	0.82	0.83	0.82	0.80
Our method	0.84	0.85	0.86	0.84

TABLE 2. Performance on ORL face database. (Gaussian kernel for KPCA)(Gaussian kernel 1 denotes $\sigma^2 = 1 \times 10^7$; Gaussian kernel 2: $\sigma^2 = 1 \times 10^8$; Gaussian kernel 3: $\sigma^2 = 1 \times 10^9$; Gaussian kernel 4: $\sigma^2 = 1 \times 10^{10}$)

Methods	Gaussian kernel 1	Gaussian kernel 2	Gaussian kernel 3	Gaussian kernel 4
Fusion method[4]	0.75	0.84	0.84	0.80
Our method	0.80	0.85	0.87	0.85

TABLE 3. Performance on Yale face database. (Polynomial kernel for KPCA)

Methods	d=2	d=3	d=4	d=5
Fusion method[4]	0.85	0.84	0.86	0.86
Our method	0.89	0.88	0.89	0.89

TABLE 4. Performance on Yale face database. (Gaussian kernel for KPCA) (Gaussian kernel 1 denotes $\sigma^2 = 1 \times 10^5$; Gaussian kernel 2: $\sigma^2 = 1 \times 10^6$; Gaussian kernel 3: $\sigma^2 = 1 \times 10^7$; Gaussian kernel 4: $\sigma^2 = 1 \times 10^8$)

Methods	Gaussian kernel 1	Gaussian kernel 2	Gaussian kernel 3	Gaussian kernel 4
Fusion method[4]	0.74	0.83	0.82	0.810
Our method	0.77	0.85	0.84	0.84

The results shows the proposed algorithm outperforms the classical method on classification performance. Our work only pays attention to the classification problem based on kernel learning. So the experiments show the classification performance, and the criterion of kernel optimization is created by increasing the classification performance. So, the kernel optimization criterion is not adaptive to clustering. The clustering application based kernel optimization is our future research work.

5. **Conclusions.** A novel parallel feature fusion based maximum margin criterion, namely discriminant parallel feature fusion, for pattern classification in this paper. We create a constrained optimization problem based on maximum margin criterion to find the fusion coefficients of the parallel feature fusion and transform the optimization problem into an eigenvalue problem, which brings two advantages, i.e., fused data has the largest class discriminant and fusion coefficients are unique. This paper only proposes a linear combination fusion strategy based on discriminant analysis, and we expect to give a promising idea for other fusion strategy. On the computation efficiency, the iteration solution method will cost much time, but this training step can be implemented off-line.

Acknowledgment. This work is supported by Suzhou Science and Technology Development Project (Grand No. syg201249 , titled by The design of elevator safety monitoring system based on Wireless Sensor Network).

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, M. H. Hayes, and D. J. Kriegman, Linear Subspace for Illumination Robust Face Recognition,*Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 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, 2005.

[4] Q. Liu, H. Lu, and S. 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, P. E. Lpez de Teruel, Nonlinear kernel-based statistical pattern analysis,*IEEE Trans. Neural Networks*, vol. 12, 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, 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, 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, pp.572–1581, 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, Jan. 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, L.T. Chen, A Survey of Video Enhancement Techniques,*Journal of Information Hiding and Multimedia Signal Processing*, vol. 3, no. 1, pp. 71-99, January 2012.
[13] Y. b. Rao, 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] A. Serrano, I. Martin de Diego, C. Conde, and E. Cabello, Recent advances in face biometrics with Gabor wavelets: A review,*Pattern Recognition Letters*, vol. 31, pp.372–381, 2010.
[15] M. Parviz, 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.
[16] Taropa, E. Srini, V.P, W. J. Lee, and T. D. Han, Data Fusion Applied on Autonomous Ground Vehicles,*The 8th International Conference Advanced Communication Technology*, vol. 1, pp.744C749, 2006.
[17] A.H. Gunatilaka, B.A. Baertlein, Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection,*IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp.577C589, 2001.
[18] L. D, Goubran, R. A, Dansereau, R.M , Robust joint audio-video talker localization in video conferencing using reliability information-II: Bayesian network fusion,*IEEE Trans. Instrumentation and Measurement*, vol. 54, pp. 1541C1547,2005.
[19] C. J. Liu, Wechsler, A shape-and texture-based enhanced Fisher classifier for face recognition,*IEEE Trans. Image Processing*, vol. 10, no. 4, pp. 598-608, 2001.
[20] X. H. Zhang, An information model and method of feature fusion,*Int. Conf. Signal Process*, vol. 2, pp. 1389C1392, 1998.
[21] R. Battiti, Using mutual information for selecting features in supervised neural net learning,*IEEE Trans. Neural Network*, vol. 5, no. 4, pp. 537C550, 1994.
[22] Haifeng Li, Tao Jiang, and Keshu Zhang, Efficient and Robust Feature Extraction by Maximum Margin Criterion,*IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 157-165, 2006.
[23] 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.
[24] F. Samaria, A. Harter, Parameterisation of a Stochastic Model for Human Face Identification,*Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, December 1994.
[25] N. Subrahmanya, Y.C. Shin, Sparse Multiple Kernel Learning for Signal Processing Applications,*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp.788-798, 2010.
[26] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, Large Scale Multiple Kernel Learning,*Journal of Machine Learning Research*, vol. 7. pp. 1531-1565, 2006.
[27] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, lp-Norm Multiple Kernel Learning,*Journal of Machine Learning Research*, vol. 12, pp.953-997, 2011.
[28] X. D. Xie, B. H. Li and X. D. Chai, Kernel-Based Nonparametric Fisher Classifier for Hyperspectral Remote Sensing Imagery,*Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 3, pp. 591-599, 2015.