Neural Networks 22 (2009) 977–987

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A novel kernel-based maximum a posteriori classification method

Zenglin Xu [a], Kaizhu Huang [b], Jianke Zhu [a], Irwin King [a,*], Michael R. Lyu [a]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

[b] Department of Engineering Mathematics, University Of Bristol, Bristol, BS8 1TR, United Kingdom

ARTICLE INFO

Article history:
Received 19 March 2008
Revised and accepted 20 November 2008

Keywords:
Kernel methods
Maximum a posteriori
Discriminant analysis

ABSTRACT

Kernel methods have been widely used in pattern recognition. Many kernel classifiers such as Support Vector Machines (SVM) assume that data can be separated by a hyperplane in the kernel-induced feature space. These methods do not consider the data distribution and are difficult to output the probabilities or confidences for classification. This paper proposes a novel Kernel-based Maximum A Posteriori (KMAP) classification method, which makes a Gaussian distribution assumption instead of a linear separable assumption in the feature space. Robust methods are further proposed to estimate the probability densities, and the kernel trick is utilized to calculate our model. The model is theoretically and empirically important in the sense that: (1) it presents a more generalized classification model than other kernel-based algorithms, e.g., Kernel Fisher Discriminant Analysis (KFDA); (2) it can output probability or confidence for classification, therefore providing potential for reasoning under uncertainty; and (3) multi-way classification is as straightforward as binary classification in this model, because only probability calculation is involved and no one-against-one or one-against-others voting is needed. Moreover, we conduct an extensive experimental comparison with state-of-the-art classification methods, such as SVM and KFDA, on both eight UCI benchmark data sets and three face data sets. The results demonstrate that KMAP achieves very promising performance against other models.

1. Introduction

Kernel methods play an important role in machine learning and pattern recognition (Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004). They have achieved success in almost all traditional tasks of machine learning, i.e., supervised learning (Mika, Ratsch, Weston, Scholkopf, & Muller, 1999; Vapnik, 1998), unsupervised learning (Schölkopf, Smola, & Müller, 1998), and semi-supervised learning (Chapelle, Schölkopf, & Zien, 2006; Xu, Jin, Zhu, King, & Lyu, 2008; Xu, Zhu, Lyu, & King, 2007; Zhu, Kandola, Ghahramani, & Lafferty, 2005). We focus here on kernel methods for supervised learning, where the basic idea is to use the so-called kernel trick to implicitly map the data from the ordinal input space to a high dimensional feature space, in order to make the data more separable. Usually, the aim of kernel-based classifiers is to find an optimal linear decision function in the feature space, based on certain criteria. The optimal linear decision hyperplane could be, for example, the one that can maximize the margin between two different classes of data (as used in the Support Vector Machine (SVM) (Vapnik, 1998)), or the one that minimizes the within-class covariance and at the same time maximizes the between-class covariance (as used in the Kernel Fisher Discriminant Analysis (KFDA) (Mika et al., 1999, 2003)), or the one that minimizes the worst-case accuracy bound (as used in the Minimax Probability Machine (Huang, Yang, King, & Lyu, 2004; Huang, Yang, King, Lyu, & Chan, 2004; Huang, Yang, Lyu, & King, 2008; Lanckriet, Ghaoui, Bhattacharyya, & Jordan, 2002)).

These kernel methods usually achieve higher prediction accuracy than their linear forms (Schölkopf & Smola, 2002). The reason is that the linear discriminant functions in the feature space can represent complex separating surfaces when mapped back to the original input space. However, one drawback of standard SVM is that it does not consider the data distribution and cannot properly output the probabilities or confidences for the resultant classification (Platt, 1999; Wu, Lin, & Weng, 2004). One needs special transformation in order to output probabilities. Therefore, it takes a lot of extra effort in order to be applied in systems that contain inherent uncertainty. In addition, the linear discriminant function can only separate two classes. For multi-category problems, we may resort to approaches such as one-against-one or one-against-others to vote on which class should be assigned (Hsu & Lin, 2002).

One approach to obtaining classification probabilities is to use a statistical pattern recognition technique, in which the probability

* Corresponding author. Tel.: +852 2609 8398; fax: +852 2603 5024.
E-mail addresses: zlxu@cse.cuhk.edu.hk (Z. Xu), k.huang@bristol.ac.uk (K. Huang), jkzhu@cse.cuhk.edu.hk (J. Zhu), king@cse.cuhk.edu.hk (I. King), lyu@cse.cuhk.edu.hk (M.R. Lyu).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.11.005

density function can be derived from the data. Future items of data can then be classified using a Maximum A Posteriori (MAP) method (Duda, Hart, & Stork, 2000). One typical probability estimation method is to assume multivariate normal density functions over the data. The multivariate normal density functions are easy to handle; moreover some problems can also be regarded as Gaussian problems if there are enough examples, although in practice the Gaussian distribution cannot be easily satisfied in the input space.

To solve these problems, in this paper we propose a Kernel-based Maximum A Posteriori (KMAP) classification method under a Gaussianity assumption in the feature space. With this assumption, we derive a non-linear discriminant function in the feature space, in contrast to current kernel-based discriminant methods that rely only on using an assumption of linear separability for the data. Moreover, the derived decision function can output the probabilities or confidences. In addition, the distribution can be very complex in the original input space when it is mapped back from the feature space. This is analogous to the case in which a hyperplane derived with KFDA or SVM in the feature space could lead to a complex surface in the input space. Therefore, this approach sets a more valid foundation than the traditional multivariate probability estimation methods that are usually conducted in the input space.

Generally speaking, distributions other than the Gaussian function can also be assumed in the feature space. However, under a distribution with a complex form, it is hard to get a closed-form solution and easy to over-fit. More importantly, with the Gaussian assumption, a kernelized version can be derived without knowing the explicit form of the mapping functions for our model, while it is still difficult to formulate the kernel version for other complex distributions.

It is important to relate our proposed model to other probabilistic kernel methods. Kernel-based exponential methods (Canua & Smola, 2006) use parametric exponential families to explicitly build mapping functions from the input space to the feature space. It is also interesting to discuss the Kernel Logistic Regression (KLR) (Zhu & Hastie, 2005), which employs the logistic regression to estimate the density function and still leads to a linear function in the kernel-induced feature space. The kernel-embedded Gaussian mixture model in Wang, Lee and Zhang (2003) is related to our model in that a similar distribution is assumed, but their model is restricted to clustering and cannot be directly used in classification.

The appealing features of KMAP are summarized as follows. First, one important feature of KMAP is that it can be regarded as a more generalized classification model than KFDA and other kernel-based algorithms. KMAP provides a rich class of family of kernel-based algorithms, based on different regularization implementations. Another important feature of KMAP is that it can output the probabilities of assigning labels to future data, which can be seen as the confidences of decisions. Therefore, the proposed method can also be seen as a Bayesian decision method, which can further be used in systems that make an inference under uncertainty (Smith, 1988). Moreover, multi-way classification is as easy as binary classification in this model because only probability calculation is involved and no one-against-one or one-against-others voting is needed. As shown in Section 2.4, KMAP has the time complexity $\mathcal{O}(n^3)$ (where n is the cardinality of data), which is in the same order as that of KFDA. In addition, the decision function enjoys the property of sparsity: only a small number of eigenvectors are needed for future prediction. This leads to low storage complexity.

The proposed algorithm can be applied in many pattern recognition tasks, e.g., face recognition, character recognition, and others. In order to evaluate the performance of our proposed method, extensive experiments are performed on eight benchmark data sets from the UCI repository and on three standard face data sets. Experimental results show that our proposed method achieves very competitive performance on UCI data. Moreover, its advantage is especially prominent in face data sets, where only a small amount of training data are available.

The remainder of this paper is organized as follows. In Section 2, we derive the kernel-based MAP classification model in the feature space and discuss the parameter estimation techniques. Then the kernel calculation procedure and the theoretical connections between the KMAP model and other kernel methods are discussed. Section 3 first reports the experiments on UCI data sets against other competitive kernel methods, then evaluates our model's performance on face data sets. Section 4 draws conclusions and lists possible future research directions.

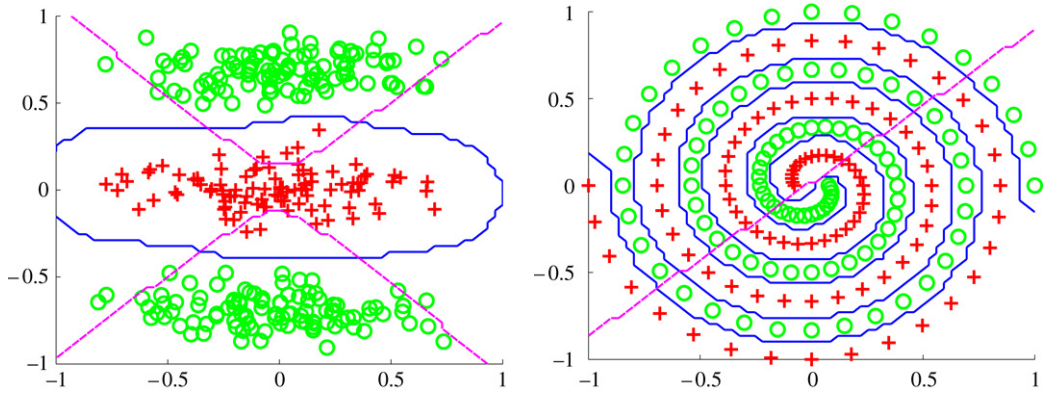
We use the following notation. Let $\mathcal{X} \in \mathbb{R}^d$ denote the original d-dimensional input space, where an instance $\mathbf{x}$ is generated from an unknown distribution. Let $\mathbb{C} = \{1, 2, \ldots, m\}$ be the set of labels where m is the number of classes. Let $P(C_i)$ denote the prior probability of class C_i. Let n_i be the number of observed data points in class C_i and n be the amount of training data. A Mercer kernel is defined as a symmetric function κ, such that $\kappa(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, where Φ is a mapping from $\mathcal{X}$ to a feature space $\mathcal{H}$. The form of kernel function κ could be a linear kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, a Gaussian RBF kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma^2)$, or a polynomial kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^p$, for some σ and p respectively. A kernel matrix or Gram matrix $G \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix such that $G_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$. G can be further written as $[G^{(1)}, G^{(2)}, \ldots, G^{(m)}]$, where $G^{(i)}$ is an $n \times n_i$ matrix and denotes the subset of G relevant to class C_i. The covariance matrix of $G^{(i)}$ is denoted by $\Sigma_{G^{(i)}}$. We denote μ_i and Σ_i as the mean vector and covariance matrix of class C_i in the feature space, respectively. The set of eigenvalues and the set of eigenvectors belonging to Σ_i are represented as Λ_i and Ω_i. We write $p(\Phi(\mathbf{x})|C_i)$ as the probability density function of class C_i.

2. Kernel-based maximum a posteriori classification

In contrast with the assumption of traditional MAP algorithms, that the data points satisfy multivariate normal distribution in the input space, we assume that the mapped data in the high dimensional feature space follow such a distribution. This is meaningful in that the distribution can be very complex in the original input space when the Gaussian distribution is mapped back from the kernel-induced feature space. In the same sense, the decision boundary can be more complex when the quadratic decision boundary is projected into the input space.

In order to make a clear illustration of the reasonability of the Gaussian distribution in the kernel-induced feature space, two synthetic data sets, **Relevance** and **Spiral**, are used in this paper. We draw the decision boundary of discriminant functions conducted in the input space and the feature space, respectively. **Relevance** is a data set where only one dimension of the data is relevant to separate the data. **Spiral** can only be separated by highly non-linear decision boundaries. Fig. 1 plots the boundaries of the discriminant functions for the traditional MAP algorithm and the kernel-based MAP algorithm on these two data sets.

It can be observed that the MAP classifier with the Gaussian distribution assumption in the kernel-induced feature space always produces more reasonable decision boundaries. For **Relevance** data, a simple quadratic decision boundary in the input space cannot produce good prediction accuracy. However, the kernel-based MAP classifier separates these two classes of data smoothly. The difference between the boundaries of these two algorithms is especially significant for **Spiral**. This indicates that the kernel-based MAP classification algorithm can better fit the distribution of data points through the kernel trick.



(a) Relevance. (b) Spiral.

Fig. 1. The decision boundaries on **Relevance** and **Spiral**. The separating lines were obtained by projecting test data over a grid. The lines in blue (dark) and magenta (dashed) represent decision boundaries for MAP algorithms with Gaussian distribution in the feature space and those in the input space, respectively.

2.1. Model formulation

Under the Gaussian distribution assumption, the conditional density function for each class $C_i (1 \le i \le m)$ is written as:

$$p(\Phi(\mathbf{x})|C_i) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \times \exp\left\{-\frac{1}{2}(\Phi(\mathbf{x}) - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\Phi(\mathbf{x}) - \mu_i)\right\}, \quad (1)$$

where N is the dimension of the feature space and $|\Sigma_i|$ is the determinant of the covariance matrix Σ_i. It is important to note that N could be infinite for the RBF kernel function. In such case, we seek to apply the kernel trick to avoid directly computing the density function, which will be verified in Section 2.3.

Taking logs on both sides of Eq. (1) and removing the constants, we can get the Mahalanobis distance function of a data point $(\mathbf{x}_i)$ to the class center (μ_i) in the feature space when each class has the same prior probability:

$$g_i(\Phi(\mathbf{x})) = (\Phi(\mathbf{x}) - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\Phi(\mathbf{x}) - \mu_i) + \log|\Sigma_i|. \quad (2)$$

In the case that different class prior probabilities are assumed, we only need to subtract $2\log P(C_i)$ in the above equation. The intuitive meaning of the function is that a data point is more likely to be assigned to a certain class with a lower Mahalanobis distance between the data point and the class center.

We revert the Mahalanobis distance function to its original class conditional density function: $p(\Phi(\mathbf{x})|C_i) = \frac{1}{(2\pi)^{N/2}}\exp(-\frac{1}{2}g_i(\Phi(\mathbf{x})))$. According to the Bayesian Theorem, the posterior probability of class C_i is calculated by

$$P(C_i|\Phi(\mathbf{x})) = \frac{p(\Phi(\mathbf{x})|C_i)P(C_i)}{\sum_{j=1}^{m} p(\Phi(\mathbf{x})|C_j)P(C_j)}. \quad (3)$$

Based on Eq. (3), the decision rule can be formulated as below:

$$\mathbf{x} \in C_w \quad \text{if } P(C_w|\Phi(\mathbf{x})) = \max_{1\le j\le m} P(C_j|\Phi(\mathbf{x})). \quad (4)$$

This means that a test data point will be assigned to the class with the maximum of $P(C_w|\Phi(\mathbf{x}))$, i.e., the MAP. Since the MAP is calculated in the kernel-induced feature space, the output model is named as the KMAP classification.

Eq. (3) is of importance because it shows that KMAP output not only a class label, but also the probability of a data point belonging to a class. This probability can thus be seen as the confidence of classification of new data points. It can be used in statistical systems that make an inference under uncertainty (Smith, 1988). If the confidence is lower than some specified threshold, the system can refuse to make an inference. This is a distinct advantage over many kernel learning methods, including SVM, which cannot easily output these probabilities.

2.2. Parameter estimation

In order to compute the Mahalanobis distance function, the mean vector and the covariance matrix for each class must be estimated. Typically, the mean vector (μ_i) and the within-covariance matrix (Σ_i) are calculated by a maximum likelihood estimation. In the feature space, they are formulated as follows:

$$\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\Phi(\mathbf{x}_j), \quad (5)$$

$$\Sigma_i = S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\Phi(\mathbf{x}_j) - \mu_i)(\Phi(\mathbf{x}_j) - \mu_i)^{\mathrm{T}}. \quad (6)$$

Directly employing the maximum likelihood estimation S_i as the covariance matrix will generate quadratic discriminant functions in the feature space. However, the covariance estimation problem is clearly ill-posed, because the number of data points in each class is usually much smaller than the number of dimensions in the kernel-induced feature space. This problem is especially obvious in face recognition tasks. The treatment of this ill-posed problem is to introduce regularization. There are several kinds of regularization methods. One of them is to replace the individual within-covariance matrices with their average, i.e.,

$$\Sigma_i = \frac{\sum_{i=1}^{m} S_i}{m} + rI, \quad (7)$$

where I is the identity matrix and r is a regularization coefficient. This method is able to substantially reduce the number of free parameters to be estimated. Moreover, it also reduces the discriminant function between two classes to a linear one. Therefore, a linear discriminant analysis method can be obtained. We will discuss its connection to Kernel Fisher Discriminant Analysis (KFDA) in Section 2.4.

Alternatively, we can estimate the covariance matrix by combining the above linear discriminant function with the quadratic one. Instead of estimating the covariance matrix in the input space (Friedman, 1989), we can apply this method in the kernel-induced feature space. After the data are centered (see Schölkopf et al. (1998) for centering data), the formulation in the feature space is as follows:

$$\Sigma_i = (1-\eta)\tilde{\Sigma}_i + \eta\frac{trace(\tilde{\Sigma}_i)}{n}I, \tag{8}$$

where

$$\tilde{\Sigma}_i = (1-\theta)S_i + \theta\tilde{S}, \tag{9}$$

$$\tilde{S} = \frac{1}{n}\sum_{l=1}^{n}\Phi(\mathbf{x}_l)\Phi(\mathbf{x}_l)^{\mathrm{T}}. \tag{10}$$

In the equations, θ ($0 \le \theta \le 1$) is a coefficient linked with the linear and quadratic discriminant term. Also, η ($0 \le \eta \le 1$) determines the shrinkage to a multiple of the identity matrix. Note that the formulation of Eq. (10) differs from the one in Friedman (1989), where $S = \sum_{i=1}^{m} S_i$. This is because it is more accurate to estimate the covariance from all samples rather than only from those belonging to a single class. The effect is particularly significant in face recognition, where the sample size is relatively small and the dimensionality of the feature space is quite high. Because of this, our approach is more capable of adjusting the effect of the regularization.

Remark. Other regularization methods can also be employed for estimating the covariance matrices. The criteria of selecting the regularization are based on specific applications of KMAP. For example, when the number of training samples is small, it is better to use the regularization method based on Eq. (8).

2.3. Kernel calculation

It is critical to represent the above formulations in a kernelized or inner product form. In the following, we demonstrate how the KMAP formulations can be kernelized without knowing the explicit form of the mapping functions.

Obviously, Eq. (2) is poorly-posed, since we are estimating the means and covariance matrices from n samples. To avoid this problem in calculating the Mahalanobis distance function, the spectral representation of the covariance matrix, i.e., $\Sigma_i = \sum_{j=1}^{N}\Lambda_{ij}\Omega_{ij}\Omega_{ij}^{\top}$ (where Λ_{ij} and Ω_{ij} are the jth eigenvalue and eigenvector of Σ_i, respectively), is utilized instead of a direct calculation (Ruiz & Lopez-de Teruel, 2001). The small eigenvalues will, in particular, drastically degrade the performance of the function overwhelmingly, because they are underestimated due to the small number of examples. In this paper, we only estimate the k largest eigenvalues and replace each remaining eigenvalue with a nonnegative number h_i. This technique is similar to that used in Principal Component Analysis (PCA) (Jolliffe, 1986), except that the non-principal eigenvalues are replaced by a constant h_i. Thus Eq. (2) can be reformulated as follows:

$$g_i(\Phi(\mathbf{x})) = \sum_{j=1}^{k}\frac{1}{\Lambda_{ij}}[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2 + \sum_{j=k+1}^{N}\frac{1}{h_i}[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2 + \log\left(h_i^{N-k}\prod_{j=1}^{k}\Lambda_{ij}\right).$$

In the above equation, $g_i(\Phi(\mathbf{x}))$ can further be represented as follows:

$$\frac{1}{h_i}\left(\sum_{j=1}^{N}[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2 - \sum_{j=1}^{k}\left(1-\frac{h_i}{\Lambda_{ij}}\right)[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2\right).$$

We define $g_{1i}(\Phi(\mathbf{x})) = \sum_{j=1}^{N}[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2$ and $g_{2i}(\Phi(\mathbf{x})) = \sum_{j=1}^{k}(1-\frac{h_i}{\Lambda_{ij}})[\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)]^2$, such that

$$g_i(\Phi(\mathbf{x})) = \frac{1}{h_i}[g_{1i}(\Phi(\mathbf{x})) - g_{2i}(\Phi(\mathbf{x}))] + \log\left(h_i^{N-k}\prod_{j=1}^{k}\Lambda_{ij}\right).$$

In the following, we show that $g_{1i}(\Phi(\mathbf{x}))$ and $g_{2i}(\Phi(\mathbf{x}))$ can be entirely written in a kernel form. To formulate the above equations, we need to calculate the eigenvalues Λ_i and eigenvectors Ω_i. However, due to the unknown dimensionality of the feature space, Σ_i cannot be computed directly. Moreover, because of the limited number of training samples, we can only express each eigenvector as the span of all the data points, as done in Schölkopf et al. (1998). The eigenvectors are in the space spanned by all the training samples, i.e., each eigenvector Ω_{ij} can be written as a linear combination of all the training samples:

$$\Omega_{ij} = \sum_{l=1}^{n}\gamma_{ij}^{(l)}\Phi(\mathbf{x}_l) = U\gamma_{ij}, \tag{11}$$

where $\gamma_{ij} = (\gamma_{ij}^{(1)}, \gamma_{ij}^{(2)}, \ldots, \gamma_{ij}^{(n)})^{\mathrm{T}}$ is an n dimensional column vector and $U = (\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n))$.

Theorem 1. *γ_{ij} and Λ_{ij} are the eigenvector and eigenvalue of the covariance matrix $\Sigma_{G^{(i)}}$, respectively.*

The proof of Theorem 1 can be found in the Appendix. Based on Theorem 1, we can express $g_{1i}(\Phi(\mathbf{x}))$ in the kernel form:

$$\begin{aligned} g_{1i}(\Phi(\mathbf{x})) &= \sum_{j=1}^{n}\gamma_{ij}^{\mathrm{T}}U^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)U\gamma_{ij} \\ &= \sum_{j=1}^{n}\left[\gamma_{ij}^{\mathrm{T}}\left(K_{\mathbf{x}} - \frac{1}{n_i}\sum_{l=1}^{n_i}K_{\mathbf{x}_l}\right)\right]^2 \\ &= \left\|K_{\mathbf{x}} - \frac{1}{n_i}\sum_{l=1}^{n_i}K_{\mathbf{x}_l}\right\|_2^2, \end{aligned}$$

where $K_{\mathbf{x}} = \{K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x})\}^{\mathrm{T}}$.

In the same way, $g_{2i}(\Phi(\mathbf{x}))$ can be formulated as the follows:

$$g_{2i}(\Phi(\mathbf{x})) = \sum_{j=1}^{k}\left(1-\frac{h_i}{\Lambda_{ij}}\right)\Omega_{ij}^{\mathrm{T}}(\Phi(\mathbf{x})-\mu_i)(\Phi(\mathbf{x})-\mu_i)^{\mathrm{T}}\Omega_{ij}.$$

Substituting (11) into the above $g_{2i}(\Phi(\mathbf{x}))$, we have:

$$g_{2i}(\Phi(\mathbf{x})) = \sum_{j=1}^{k}\left(1-\frac{h_i}{\Lambda_{ij}}\right)\gamma_{ij}^{\mathrm{T}}\left(K_{\mathbf{x}} - \frac{1}{n_i}\sum_{j=1}^{n_i}K_{\mathbf{x}_j}\right) \times \left(K_{\mathbf{x}} - \frac{1}{n_i}\sum_{j=1}^{n_i}K_{\mathbf{x}_j}\right)^{\mathrm{T}}\gamma_{ij}.$$

Remark. In calculating $g_{2i}(\Phi(\mathbf{x}))$, only the k largest eigenvalues and relevant eigenvectors are selected for each class. In Williams and Seeger (2000) and Yang, Frangi, Yang, Zhang and Jin (2005), it is shown that the eigenvalue spectrum of the covariance matrix of the Gram matrix rapidly decays and thus is of low rank. This reinforces the theoretical basis of KMAP from another perspective.

Now, the discriminant function in the feature space $g_i(\Phi(\mathbf{x}))$ can be finally written in a kernel form, where N is substituted with the cardinality of data n.

We summarize the proposed KMAP algorithm in Fig. 2.

The overall time complexity of the algorithm is determined by Step 5 and Step 6. These steps involve computing the within-class

Algorithm 1: The KMAP Algorithm for Classification.

1. Choose a kernel function $\kappa(\mathbf{x}, \mathbf{y})$, which can be a linear kernel function, an RBF kernel or a polynomial kernel, etc.
2. Center the training data in the kernel-induced feature space.
3. Tune parameters (θ, η) and set k using training data.
4. Compute the Mahalanobis distance of each test sample to each class center $g_i(\Phi(\mathbf{x}))$ according to Eq. (11).
5. Make a decision according to the MAP rule (Eq. (4)).

Fig. 2. The KMAP algorithm for classification.

Table 1
The relationship among KMAP and other kernel methods.

Parameter setting		Kernel methods
θ	η	
0	0	A quadratic discriminant method
1	0	A linear discriminant method
1	1	The nearest mean classifier
0	1	The weighted nearest mean classifier

covariance matrix, and the complexity is $\mathcal{O}(n^2)$. In addition, it will cost $\mathcal{O}(n^3)$ operations to solve the eigenvalues and eigenvectors. Hence, KMAP has the same time complexity as KFDA. The storage complexity, which involves $\mathcal{O}(kn)$ for storing k columns of the covariance matrix, can be deduced because the value of k is much smaller than n. We will evaluate the scale of k in the experiments.

2.4. Connection with other kernel methods

The KMAP model is a generalized classification model and can be reduced to other kernel-based classification methods with different implementations of parameter estimation.

In the regularization method based on Eq. (8), by varying the settings of θ and η, other kernel-based classification methods can be derived. When $(\theta = 0, \eta = 0)$, the KMAP model represents a quadratic discriminant method in the kernel-induced feature space; when $(\theta = 1, \eta = 0)$, it represents a kernel discriminant method; and when $(\theta = 0, \eta = 1)$ or $(\theta = 1, \eta = 1)$, it represents the nearest mean classifier. Therefore, by varying θ and η, different models can be generated from different combinations of quadratic discriminant, linear discriminant and the nearest mean methods. The relationship among these kernel methods is summarized in Table 1.

We show in the following, that a special case of the regularization method when $\theta = 1$ and $\eta = 0$ will reduce to the well-known Kernel Fisher Discriminant Analysis (KFDA). If both classes are assumed to have the same covariance structure for a binary class problem (i.e., $\Sigma_i = \frac{\Sigma_1+\Sigma_2}{2}$) it leads to a linear discriminant function. Assuming all classes have the same class prior probabilities, $g_i(\Phi(\mathbf{x}))$ can be derived as:

$$\begin{aligned} g_i(\Phi(\mathbf{x})) &= (\Phi(\mathbf{x}) - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (\Phi(x) - \mu_i) \\ &= (\Phi(\mathbf{x}) - \mu_i)^{\mathrm{T}} \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\Phi(x) - \mu_i), \end{aligned}$$

where $i = 1, 2$. We can reformulate this equation in the following form: $g_i(\Phi(\mathbf{x})) = \mathbf{w}_i \Phi(\mathbf{x}) + b_i$, where

$$\begin{aligned} \mathbf{w}_i &= -4(\Sigma_1 + \Sigma_2)^{-1} \mu_i, \\ b_i &= 2\mu_i^{\mathrm{T}} (\Sigma_1 + \Sigma_2)^{-1} \mu_i. \end{aligned}$$

The decision hyperplane is $f(\Phi(\mathbf{x})) = g_1(\Phi(\mathbf{x})) - g_2(\Phi(\mathbf{x}))$, i.e.,

$$\begin{aligned} f(\Phi(\mathbf{x})) &= (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) \Phi(\mathbf{x}) \\ &\quad - \frac{1}{2} (\mu_1 - \mu_2)^{\mathrm{T}} (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 + \mu_2). \end{aligned}$$

Table 2
Overview of the experimental data sets used.

Data set	# samples	# features	# classes
Twonorm	1000	21	2
Breast	683	9	2
Ionosphere	351	34	2
Pima	768	8	2
Sonar	208	60	2
Iris	150	4	3
Wine	178	13	3
Segment	210	19	7

This equation is just the formulation of KFDA (Kim, Magnani, & Boyd, 2006; Mika et al., 1999). Therefore, KFDA can be viewed as a special case of KMAP when all classes have the same covariance structure.

Remark. KMAP thus provides a rich class of kernel-based classification algorithms using different regularization methods. This makes KMAP a flexible framework for classification adaptive to data distribution.

3. Experiments

In this section, we evaluate the proposed KMAP method on eight UCI data sets and three facial image data sets. As the classical methods and the state-of-the-art method in the face recognition task differ from tradition classification problems, we employ different comparison algorithms.

3.1. Experimental data sets

We describe these two batches of data sets for further evaluation in the following. The first batch comprises eight UCI data sets and the second comprises three facial image data sets.

3.1.1. UCI data sets

Eight data sets from the UCI machine learning repository, with different numbers of samples, features and classes, are chosen to test the performance of a variety of methods. Table 2 summarizes the information of these data sets.

3.1.2. Facial image data sets

To make comprehensive evaluations, we have collected three different kinds of data sets for our experiments. One is the Facial Recognition Technology (FERET) Database (Phillips, Moon, Rizvi, & Rauss, 2000). The second is the Face Recognition Grand Challenge (FRGC) data set (Phillips et al., 2005). The above two data sets are the de-facto standard data sets for face recognition evaluation. The third data set is the Yahoo! News facial images data set, which was obtained by crawling from the Web (Berg et al., 2004). These facial data sets are widely used for the performance evaluation of face recognition (Zhu, Hoi, & Lyu, 2008). In the following, we first describe the details of these data sets. Then we discuss our preprocessing methods for face extraction and feature representation.

FERET Face Data Set. In our experiment, 239 persons in the FERET data set are selected, and there are four gray scale 256×384 images for each individual. Among the four images, two images are from the FA/FB set, respectively, and the remaining two images are from DupI set. Therefore, there are a total of 956 images for evaluation. Since the images are acquired from different photo sessions, both the illumination conditions and the facial expressions may vary. All images are cropped and normalized by aligning the centers of the eyes to predefined positions, according to the manually located eye positions supplied by the FERET data. Fig. 3 depicts six individuals from this data set. The top two rows show the example images, the

Fig. 3. Example images from the FERET data set, cropped and normalized to the size of 128 × 128.

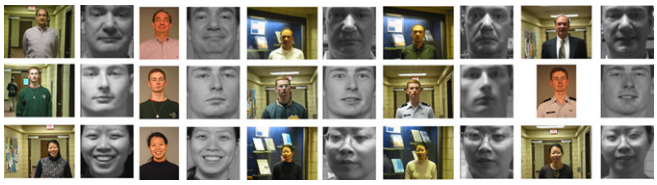


Fig. 4. Example images from the FRGC data set, cropped and normalized to the size of 128 × 128.

first row from FA, and the second one from FB; while the bottom two rows are the examples from DupI.

FRGC Data Set. The FRGC data set (Phillips et al., 2005)[1] is the current benchmark for performance evaluation of face recognition techniques. We adopt the FRGC version-1 data set (Spring 2003) for the evaluation of our face recognition method. The data set used in our experiment consists of 1920 images, corresponding to 80 individuals selected from the original collection. Each individual has 24 controlled or uncontrolled color images. The faces are automatically detected and normalized through a face detection method and an extraction method. Fig. 4 shows geometrically normalized face images cropped from the original FRGC images, with the cropped regions resized to a size of 128 × 128.

Yahoo! News Face Data Set. The Yahoo! News Face data set was constructed by Berg et al. (2004) from about half a million captioned news images collected from the Yahoo! News Web site. It consists of a large number of photographs taken in real life conditions, rather than in the controlled environments widely used in face recognition evaluation. As a result, there are a large variety of poses, illuminations, expressions, and environmental conditions. There are 1940 images, corresponding to 97 largest face clusters selected to form our experimental data set, in which each individual cluster has 20 images. As with the other data sets, faces are cropped from the selected images using the face detection and extraction methods. Only relevant face images are retained when there are multiple faces in one image. Fig. 5 presents examples selected from the Yahoo! News images and the extracted faces. All these face images are geometrically normalized.

Facial Feature Extraction. To enable an automatic face recognition scheme, we cascade a face detector (Viola & Jones, 2004) with the Active Appearance Models (AAMs) (Cootes, Edwards, & Taylor, 2001) to locate faces and facial features in the input images. The performance in terms of the correct registration is greatly dependent on the image conditions. In fact, only about 30 images failed for the FRGC data set (5660 images). Similarly, the correct registration rate for the Yahoo! News face data set was around 80%. Many effective feature extraction methods have been

[1] Accessible from http://www.frvt.org/FRGC.

Fig. 5. Example images from the Yahoo! News Face data set, cropped and normalized to the size of 128 × 128.

Table 3
Overview on the face image data sets used in the experiments.

Data set	# total	# person	# per person
FERET	956	239	4
FRGC	1920	80	24
Yahoo! News	1940	97	20

proposed to address the task, such as Local Binary Pattern (Ahonen, Hadid, & Pietikainen, 2004; Rodriguez & Marcel, 2006) and Gabor wavelets transform. Among these methods, the Gabor wavelets representation of facial image has been widely accepted as a promising approach (Liu & Wechsler, 2002). From earlier studies in the area of signal processing, Lades et al. (1993) empirically suggested that good performance can be achieved by extracting Gabor wavelet features of 5 different scales and 8 orientations. In our experiments, we employ a similar approach by applying Gabor wavelet transform on each image (scaled to 128 × 128) at 5 scales and 8 orientations. Finally, we normalize each sub-image to form a feature vector $\mathbf{x} \in \mathbf{R}^n$ with the sample scale reduced to 64, which results in a 10240-dimensional feature vector for each facial image.

In summary, the detailed statistics of the data sets used in our experiments is listed in Table 3.

3.2. Experiments on UCI data sets

In this section, we conduct experiments on eight benchmark data sets. We first implement many other competitive methods and compare them with our proposed algorithm. Then we discuss and analyze the experimental results.

3.2.1. Comparison algorithms

We provide a brief introduction to the comparison algorithms in this section. Specifically, we compare our proposed model with the Modified Quadratic Discriminant Function (Kimura, Takashina, T. S., & M. Y., 1987), KFDA, the Kernel Fisher Quadratic Discriminant Analysis (KFQDA) (Huang, Hwang, & Lin, 2005), and SVM. Due to the popularity of SVM, we only focus on introducing MQDF, KFDA, and KFQDA in the following.

In statistical pattern recognition, the probability density function can first be estimated from the data. Then future examples could be assigned to the class with the MAP. One typical probability estimation method is to assume a multivariate normal density function over the data. From the multivariate normal distribution, the Quadratic Discriminant Function (Duda et al., 2000; Fukunaga, 1990) can be derived, which achieves the minimum mean error rate under Gaussianity and is also monotonic with an increase of the feature size (Waller & Jain, 1978). In Kimura et al. (1987), a Modified Quadratic Discriminant Function (MQDF) less sensitive to the estimation error, is proposed. Friedman (1989) improves the performance of QDF by the covariance matrix interpolation.

Another type of classifier does not assume the probability density functions in advance, but is designed directly on data samples. An example is the Fisher discriminant analysis (FDA). FDA optimizes the Rayleigh coefficient to find a direction which maximizes the projected class means while minimizing the within-class variance in this direction denominator. It can be derived as a maximum likelihood method, as well as a Bayes classifier, where the data are under a Gaussian assumption. Mika et al. extend FDA to a non-linear space by the kernel trick (Mika et al., 1999, 2003), which results in the Kernel Fisher Discriminant Analysis (KFDA) model.

To supplement the statistical justification of KFDA, (Huang et al., 2005) extends the maximum likelihood method and Bayes classification to the kernel generalization under the Gaussian Hilbert space assumption. The authors of Huang et al. (2005) do not directly kernelize the quadratic forms in terms of kernel values. They instead define a Gaussian measure on a Hilbert space through joint normal distributions of inner products between the random element and an arbitrarily selected finite "coordinate system". Thus the kernel matrix is usually employed as the input data of FDA, and the derived model is named Kernel Fisher Quadratic Discriminant Analysis (KFQDA).

3.2.2. Experimental results

The parameters of different algorithms are set, as follows. In all kernel methods, a Gaussian-RBF kernel is used. The margin parameter in SVM and the kernel width in RBF kernel are all tuned by 10-cross validation. In KMAP, we use the contribution of features to the covariance matrix to determine k. Only eigenvalues satisfying $\frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j} \geq \alpha$ can be incorporated into the distance function (Eq. (2)). In KMAP, α is set to 0.001; while in MQDF, the range of k is relatively small and we select k by cross validation. PCA is used as the regularization method in KFQDA, where eigenvectors are chosen, corresponding to 99% of the total variations. The regularization parameter r is set to 0.001 in KFDA. In both KMAP and MQDF, h takes the value of λ_{k+1}. In KMAP, parameter η is set to 0.001 for all data sets and θ is searched from the range of [0.001, 1] with a step value of 0.025. All experimental results are obtained by averaging 10 runs, and each run is executed with 10-cross validation for each data set.

To analyze the effect of k in KMAP, we list the average value of k for each data set in Table 4 (where k is rounded up to its nearest integer). It can be noted that k is very small and close to the dimension of the input space for each problem. This shows that KMAP contains a sparsity property. For an m-class problem, mk vectors with larger eigenvalues in the covariance matrix need storing for a future prediction. In this sense, the meaning of these vectors is very similar to that of the support vectors in SVM. Furthermore, the storage requirement of these vectors is very small.

Table 5 reports the average prediction accuracy with the standard errors of each data set for all algorithms. It can be observed that KMAP outperforms MQDF in all data sets. These results show that the data become more separable after the mapping to the high dimensional feature space. This is because that Gaussianity is easier to satisfy in a specially chosen feature space, and it can fit distributions that are very complex in the original input space. Compared with other kernel classifiers, KMAP also achieves better results on a number of data sets. When the problem better fits the Gaussian distribution, the advantages of KMAP are especially pronounced. The reason is that KMAP accurately captures the prior distribution of the problems and incorporates it into the discriminant function.

3.3. Experiments on facial image data sets

In this section, we report empirical evaluations of the KMAP algorithm for face recognition. For performance comparison, we also implement several other approaches for face recognition, i.e., LDA (Belhumeur, Hespanha, & Kriegman, 1997), Kernel PCA (Schölkopf et al., 1998), SVM, and Regularized Kernel Fisher Discriminant (RKFD)[2] (Liu, 2006). LDA and Kernel PCA are two conventional methods for face recognition tasks. RKFD is currently regarded as the state-of-the-art approach. In the following sections, we first introduce the comparison algorithms. Then we present and discuss our experimental results.

3.3.1. Comparison algorithms

Various comparison algorithms are introduced in this section. In the following, we introduce the competitive algorithms (i.e., LDA, RKFD, and KPCA) in the context of face recognition. Due to the popularity of SVM, we shall avoid introducing SVM.

In face recognition, the number of training samples for each face is usually very small. This is known as the Small Sample Size (S3) problem. In order to deal with this problem, direct Linear Discriminant Analysis (LDA) methods (Yu & Yang, 2001) were presented to process the data in the high-dimensional input space. Moreover, along with the typical Fisher's discriminant criterion $J = \mathbf{tr}(\Sigma_w^{-1}\Sigma_b)$, where Σ_w is the within-class covariance matrix and Σ_b is the between-class covariance matrix, several methods based on variants of this criterion (see for example Wang, Plataniotis, Lu, and Venetsanopoulos (2006) and Zheng, Zhao, and Zou (2004)) were employed to deal with the S3 problem in the high dimensional input space. These methods eventually come down to a trace quotient problem (Yan & Tang, 2006), and can be solved by either generalized eigen-decomposition or diagonalization.

In addition, kernel methods have been widely used in face recognition. KPCA projects face images to directions with the greatest variances. Kernel Fisher Discriminant Analysis (Mika et al., 1999) has been suggested to solve the problem of Fisher's linear discriminant in the kernel-induced feature space, thereby yielding a nonlinear discriminant in the input space. Compared with other kernel methods such as SVMs, KFDA enjoys better outputs with probabilistic interpretations and natural solutions to multi-class classification problems. Face recognition algorithms based on KFDA include Generalized Discriminant Analysis (GDA) (Baudat & Anouar, 2000; Howland, Wang, & Park, 2006), Kernel Fisher Faces (Yang, 2002), Kernel Direct Discriminant Analysis (KDDA) (Lu, Plataniotis, & Venetsanopoulos, 2003), Regularized Kernel Fisher Discriminant (Liu, 2006) and other variants (Aladjem, 1998; Chen, Yuen, Huang, & Dai, 2005; Dai & Yuen, 2007; Yang et al., 2005). Among these approaches, RKFD is currently regarded as the state-of-the-art approach, and has gained much success in the recent FRGC contests.

On the other hand, probabilistic models are also widely used in face recognition (Liu & Wechsler, 1998; Moghaddam & Pentland, 1997; Zhou & Chellappa, 2006). One approach is to assume a Gaussian distribution in the original image space (Moghaddam & Pentland, 1997) or the PCA-projected space (Liu & Wechsler, 1998). In Moghaddam and Pentland (1997), a Gaussian distribution is first assumed in the original image space and then the posterior density function is estimated through the primary projections produced by PCA. Other probabilistic reasoning models based on the assumption of Gaussian distribution in the PCA-projected space are proposed in Liu and Wechsler (1998). The posterior density functions of those models are constructed by the diagonal

[2] It is named as Kernel Fisher Analysis (KFA) in the original paper.

Table 4
The value of k for all data sets.

Data set	Twonorm	Breast	Iono	Pima	Sonar	Iris	Wine	Segment
k value	22	14	11	21	82	5	16	11

Table 5
The prediction results of KMAP and other methods.

Data set (%)	KMAP	SVM	MQDF	KFDA	KFQDA
Iono	**94.6** ± 0.5	94.1 ± 0.7	89.6 ± 0.5	94.2 ± 0.1	93.6 ± 0.4
Breast	**96.8** ± 0.2	96.5 ± 0.4	96.5 ± 0.1	96.4 ± 0.1	96.5 ± 0.1
Twonorm	**97.6** ± 0.7	96.1 ± 0.4	97.4 ± 0.4	96.7 ± 0.5	97.3 ± 0.5
Sonar	**90.4** ± 0.7	86.6 ± 1.0	83.7 ± 0.7	88.3 ± 0.3	85.1 ± 1.9
Pima	75.9 ± 0.4	**77.9** ± 0.7	73.1 ± 0.4	71.0 ± 0.5	74.1 ± 0.5
Iris	**97.3** ± 0.4	96.2 ± 0.4	96.0 ± 0.1	95.7 ± 0.1	96.8 ± 0.2
Wine	**99.4** ± 0.2	98.8 ± 0.1	99.2 ± 1.3	99.1 ± 0.1	96.9 ± 0.7
Segment	89.7 ± 1.0	**92.8** ± 0.7	86.9 ± 1.2	91.6 ± 0.3	85.8 ± 0.8
Average	**92.71**	92.38	90.30	91.63	90.76

within-covariance matrix or by the average of each within-covariance matrix. However, discriminant information may be discarded during the PCA reduction. In practice, the performance of these two PCA-related methods is not better than that of the kernel discriminant-based approaches. In addition, the Gaussian distribution is not easily satisfied in either the original image space or the PCA-projected space.

3.3.2. Experimental results

The LDA algorithm is employed as the baseline method for evaluating the proposed face recognition approach. The implemented baseline method[3] is similar to the Fisherfaces method (Belhumeur et al., 1997; Yang & Yang, 2003), which applies LDA after PCA dimensionality reduction. Note that the LDA baseline algorithm is performed on the normalized intensities, while the Kernel PCA, SVM, RKFD and the proposed approach are performed on the Gabor wavelet features.

We introduce the following experimental protocol for all the tests. Each data set is partitioned into a training set and a testing set. In practice, the number of training faces for each person is usually small, while face variation for a certain person may be large in the real testing or recognition stage (Liu, 2006). To be consistent with reality, we investigate the performance when the number of training samples is small. We increase the amount of training data in each class gradually (less than 2, 5, 5 respectively for the three data sets), while the remaining data are employed as testing data. A variation of ten-fold cross validation is performed in the experiments. For each test, the training data points are randomly selected. The linear kernel function is used to compute the kernels. For selecting eigenspaces of KPCA and RKFD, we choose the eigenvectors corresponding to 99% of the total variations. We use the LibSVM[4] toolbox to implement SVM. Due to the small training sample size, the trade-off parameter C of SVM is set according to the default settings of LibSVM. In KMAP, parameter η is set to 0.01 for all data sets, and θ is searched within the range [0.001, 1] with a step value of 0.025.

For the FERET data set, Table 6 summarizes the experimental results with different settings, where L denotes the amount of training data in each class. We can first observe that the proposed KMAP method achieves significant improvements over other methods when the training sample size for each class is equal to one. Moreover, RKFD performs similarly to KPCA, because RKFD is reduced to a kind of regularized KPCA method, with only one sample for each class. When the training sample size is two, RKFD, SVM and KMAP evidently outperform the baseline method, and KPCA, and RKFD performs slightly better than KMAP and SVM.

Table 6
Correct recognition rates with standard deviation on the FERET data set.

L (%)	1	2
LDA	25.69 ± 2.54	55.64 ± 0.55
KPCA	42.96 ± 1.38	55.09 ± 1.60
RKFD	43.02 ± 1.28	**92.25** ± 1.28
KMAP	**84.10** ± 0.98	91.28 ± 1.22
SVM	80.36 ± 0.96	92.05 ± 0.93

In the FRGC data set, similar evaluations are conducted. Table 7 shows the experimental results of overall correct recognition rates. We can find that the recognition performance of KMAP is better than other algorithms in many cases. In comparison with the RKFD method, the improvements are particularly more significant when the amount of training data in each class is small. For example, the correct recognition rate in KMAP is more than twice of that in RKFD when the amount of training data in each class equals one. Furthermore, KMAP outperforms RKFD by over 24% when the sample size in each class equals two. We also find that SVM, which is usually regarded as the state-of-the art approach in dealing with the small sample scenario, performs no better than our proposed KMAP, in many cases. This indicates the effectiveness of KMAP for small sample problems. Quantitatively, the displayed results of KMAP in Table 5 is a little different from the results in the preliminary version (Xu, Huang, Zhu, King, & Lyu, 2007), where a different parameter estimation method is used. But they are qualitatively consistent.

Table 8 depicts the evaluation results on the Yahoo! News image data set. The results indicate that KMAP performs better than the reference methods in almost all cases. We find that the face recognition task on Web images is more challenging than the standard evaluation sets, such as FERET and FRGC. This is because the images collected from the Yahoo! News image data set inherit more variations of pose illumination conditions.

From the above experimental results on three data sets, we can conclude that our proposed KMAP approach is more effective than the RKFD method and other methods for face recognition, when dealing with the small sample size problem. This is a critical property of KMAP for practical applications.

4. Conclusion and future work

In this paper, we present a novel kernel classifier named Kernel-based Maximum A Posteriori, which assumes multivariate

[3] A regularization term is added into the LDA optimization $(S_b+\gamma I)^{-1}S_w$ in order to ensure numerical stability, where $\gamma = 0.001$. In addition, the Euclidean distance is employed as similarity measurement.

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Table 7
Correct recognition rates with standard deviation on the FRGC data set.

L (%)	1	2	3	4	5
LDA	13.30 ± 4.79	20.31 ± 6.61	34.12 ± 2.84	36.79 ± 3.15	43.94 ± 3.95
KPCA	15.23 ± 0.89	18.37 ± 1.74	21.69 ± 1.52	23.70 ± 1.84	26.05 ± 1.89
RKFD	15.32 ± 0.87	35.72 ± 2.08	52.15 ± 2.15	62.24 ± 1.30	71.00 ± 1.05
KMAP	**34.46** ± 1.86	44.41 ± 1.47	**59.33** ± 1.90	**65.68** ± 1.29	**71.92** ± 1.16
SVM	30.52 ± 1.48	**47.66** ± 0.62	57.23 ± 2.46	64.55 ± 1.25	69.43 ± 1.78

Table 8
Correct recognition rates with standard deviation on the Yahoo! Web News data set.

L (%)	1	2	3	4	5
LDA	8.35 ± 1.15	13.04 ± 1.48	16.11 ± 1.17	17.99 ± 2.28	19.85 ± 1.68
KPCA	9.37 ± 0.41	13.76 ± 1.04	16.62 ± 1.49	19.48 ± 1.38	20.94 ± 0.10
RKFD	9.40 ± 0.40	22.91 ± 0.47	33.78 ± 1.08	40.83 ± 0.52	47.08 ± 1.09
KMAP	16.10 ± 1.65	**29.15** ± 1.32	**36.51** ± 1.67	**44.16** ± 1.05	**48.13** ± 0.65
SVM	**17.21** ± 0.82	26.62 ± 1.80	35.88 ± 1.62	41.07 ± 0.98	46.80 ± 0.50

normal distribution in the high dimensional feature space and builds a quadratic discriminant function in the feature space. The theoretical reasonability of KMAP lies in the fact that the Gaussian distribution in the kernel-induced feature space can fit distributions that are very complex in the original input space, and that KMAP is robust to estimation errors. Compared with the state-of-the-art classifiers, the advantages of KMAP include that the prior information of distribution is incorporated, and it naturally outputs probability or confidence in making a decision. KMAP can easily handle multi-category problems, and no schemes such as one-against-one or one-against-all are needed. Furthermore, we conduct an extensive experimental comparison with several state-of-the-art classification methods on a number of UCI data sets and face image data sets. The experimental results show that KMAP achieves very promising performance against other models.

Our future research directions mainly include finding more precise estimations of parameters and reducing the time complexity. The parameters k and h in this paper are set experimentally. However, it would be desirable to choose them theoretically so that the prediction error can be bounded. In addition, it is currently still very hard to theoretically or empirically evaluate the Gaussian distribution in the feature space, due to the unknown dimensionality of the feature space. We leave this as an open problem. Furthermore, we may consider incorporating unlabeled information into the proposed scheme to design semi-supervised learning approaches, and apply the proposed approach to other applications.

Acknowledgments

The work described in this paper was substantially supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK4150/07E and Project No. CUHK4125/07).

Appendix

Proof of Theorem 1. We prove this theorem using the maximum likelihood estimation of the covariance matrix. For other estimations, the theorem can be similarly proved. According to the definition of covariance matrix, a pair of eigenvalue and eigenvector $\{\Lambda_{ij}, \Omega_{ij}\}$ must satisfy $\Sigma_i \Omega_{ij} = \Lambda_{ij}\Omega_{ij}$, or equivalently,

$$\Omega_{ij}^{\mathrm{T}} \Sigma_i \Omega_{ij} = \Lambda_{ij}, \tag{12}$$

since the eigenvectors are orthogonal to each other.

If substituting Eqs. (6) and (11) into Eq. (12), we can have

$$\gamma_{ij}^{\mathrm{T}} U^{\mathrm{T}} \frac{1}{n_i} \sum_{j=1}^{n_i} (\Phi(\mathbf{x}_j) - \mu_i)(\Phi(\mathbf{x}_j) - \mu_i)^{\mathrm{T}} U \gamma_{ij} = \Lambda_{ij}.$$

It can be reformulated as

$$\gamma_{ij}^{\mathrm{T}} M \gamma_{ij} = \Lambda_{ij}, \tag{13}$$

where $M = \frac{1}{n_i} \sum_{j=1}^{n_i} (K_{\mathbf{x}_j} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l})(K_{\mathbf{x}_j} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l})^{\mathrm{T}}$.

By the maximum likelihood estimation of the covariance matrix $\Sigma_{G^{(i)}}$, we have

$$\Sigma_{G^{(i)}} = \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\left(K_{\mathbf{x}_j} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right) \left(K_{\mathbf{x}_j} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right)^{\mathrm{T}} \right].$$

Note that this is equal to M. Moreover, $\Sigma_{G^{(i)}} \in \mathbb{R}^{n \times n}$, thus j is restricted to the range of $[1, n]$. Therefore, γ_{ij} and Λ_{ij} are actually the respective eigenvector and eigenvalue of $\Sigma_{G^{(i)}}$. □

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2004). Face recognition with local binary patterns. In *Proceedings of 2004 European conference on computer vision* (pp. 469–481). *Vol. 1*.

Aladjem, M. (1998). Nonparametric discriminant analysis via recursive optimization of Patrick–Fisher distance. *IEEE Transactions on Systems, Man and Cybernetics*, *28*(2), 292–299.

Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, *12*(10), 2385–2404.

Belhumeur, P. N., Hespanha, J., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., & Yee, W. T. et al. (2004). Names and faces in the news. In *Proceedings of 2004 IEEE computer society conference on computer vision and pattern recognition* (pp. 848–854). *Vol. 2*.

Canua, S., & Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing*, *69*, 714–720.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.) (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Chen, W.-S., Yuen, P., Huang, J., & Dai, D.-Q. (2005). Kernel machine-based one-parameter regularized Fisher discriminant method for face recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *35*(4), 659–669.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685.

Dai, D.-Q., & Yuen, P. C. (2007). Regularized discriminant analysis and its application to face recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *37*(4), 1080–1085.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley-Interscience Publication.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of American Statistics Association*, *84*(405), 165–175.

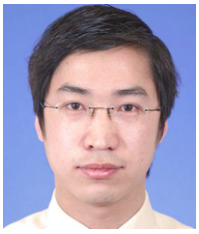
Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). San Diego: Academic Press.

Howland, P., Wang, J., & Park, H. (2006). Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, *39*(2).

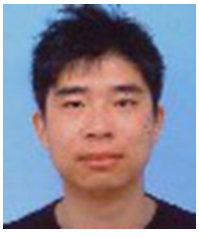
Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*, 415–425.

Huang, K., Yang, H., King, I., & Lyu, M. R. (2004). Learning large margin classifiers locally and globally. In R. Greiner, & D. Schuurmans (Eds.), *Proceedings of the twenty-first international conference on machine learning* (pp. 401–408).
Huang, K., Yang, H., King, I., Lyu, M. R., & Chan, L. (2004). Minimum error minimax probability machine. *Journal of Machine Learning Research*, *5*, 1253–1286.
Huang, K., Yang, H., Lyu, M. R., & King, I. (2008). Maxi–min margin machine: Learning large margin classifiers localy and globally. *IEEE Transactions on Neural Networks*, *19*(2), 260–272.
Huang, S. -Y., Hwang, C. -R., & Lin, M. -H. (2005). *Kernel Fisher's discriminant analysis in Gaussian reproducing kernel Hilbert space. Technical report*. Academia Sinica, Taiwan, ROC.
Jolliffe, I. (1986). *Principal component analysis*. New York: Springer-Verlag.
Kim, S.-J., Magnani, A., & Boyd, S. (2006). Optimal kernel selection in kernel Fisher discriminant analysis. In *ICML '06: Proceedings of the 23rd international conference on Machine learning* (pp. 465–472). New York, NY, USA: ACM Press.
Kimura, F., Takashina, K., S., T., & Y., M. (1987). Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*, 149–153.
Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*(5), 300–311.
Lanckriet, G. R. G., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*, 555–582.
Liu, C. (2006). Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 725–737.
Liu, C., & Wechsler, H. (1998). Probabilistic reasoning models for face recognition. In *Proceedings of 1998 IEEE computer society conference on computer vision and pattern recognition* (pp. 827–832).
Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, *11*, 467–476.
Lu, J., Plataniotis, K., & Venetsanopoulos, A. (2003). Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, *14*(1), 117–126.
Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Muller, K. (1999). Fisher discriminant analysis with kernels. In *Proceedings of IEEE neural network for signal processing workshop* (pp. 41–48).
Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A., & Müller, K. (2003). Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(5), 623–633.
Moghaddam, B., & Pentland, A. P. (1997). Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 696–710.
Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., & Hoffman, K. et al. (2005). Overview of the face recognition grand challenge. In *Proceedings of 2005 IEEE computer society conference on computer vision and pattern recognition* (pp. 947–954). *Vol. 1*.
Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 1090–1104.
Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.
Rodriguez, Y., & Marcel, S. (2006). Face authentication using adapted local binary pattern histograms. In *Proceedings of 2006 European conference on computer vision* (pp. 321–332). *Vol. 4*.
Ruiz, A., & Lopez-de Teruel, P. (2001). Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, *12*(1), 16–32.
Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.
Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.
Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. UK: Cambridge University Press.
Smith, J. (1988). *Decision analysis: A Bayesian approach*. Chapman and Hall.
Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.
Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.
Waller, W., & Jain, A. (1978). On the monotonicity of the performance of Bayesian classifiers. *IEEE Transactions on Information Theory*, *24*, 392–394.
Wang, J., Lee, J., & Zhang, C. (2003). Kernel trick embedded Gaussian mixture model. In *Proceedings of the 14th international conference on algorithmic learning theory* (pp. 159–174).
Wang, J., Plataniotis, K., Lu, J., & Venetsanopoulos, A. (2006). On solving the face recognition problem with one training sample per subject. *Pattern Recognition*, *39*(9), 1746–1762.
Williams, C. K. I., & Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. In *Proceedings of 17th international conf. on machine learning* (pp. 1159–1166). San Francisco, CA: Morgan Kaufmann.
Wu, T., Lin, C., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*, 975–1005.
Xu, Z., Huang, K., Zhu, J., King, I., & Lyu, M.R. (2007). Kernel maximum a posteriori classification with error bound analysis. In *ICONIP'07: Proceedings of international conference on neural information processing* (pp. 841–850).
Xu, Z., Jin, R., Zhu, J., King, I., & Lyu, M. R. (2008). Efficient convex relaxation for transductive support vector machine. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1641–1648). Cambridge, MA: MIT Press.
Xu, Z., Zhu, J., Lyu, M.R., & King, I. (2007). Maximum margin based semi-supervised spectral kernel learning. In *IJCNN'07: Proceedings of 20th international joint conference on neural network* (pp. 418–423).
Yan, S., & Tang, X. (2006). Trace quotient problems revisited. In *Proceedings of 2006 European conference on computer vision* (pp. 232–244). *Vol. 2*.
Yang, J., Frangi, A. F., Yang, J., Zhang, D., & Jin, Z. (2005). Kpca plus lda: A complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(2), 230–244.
Yang, J., & Yang, J.-Y. (2003). Why can LDA be performed in PCA transformed space. *Pattern Recognition*, *36*(2), 563–566.
Yang, M.-H. (2002). Kernel eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods. In *FGR '02: Proceedings of the fifth IEEE international conference on automatic face and gesture recognition* (p. 215). Washington, DC, USA: IEEE Computer Society.
Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data with application to face recognition pattern recognition. *Pattern Recognition*, *34*(10), 2067–2070.
Zheng, W., Zhao, L., & Zou, C. (2004). An efficient algorithm to solve the small sample size problem for LDA. *Pattern Recognition*, *37*(5), 1077–1079.
Zhou, S. K., & Chellappa, R. (2006). From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(6), 917–929.
Zhu, J., & Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, *14*(21), 185–205.
Zhu, J., Hoi, S., & Lyu, M. R. (2008). Face annotation using transductive kernel Fisher discriminant. *IEEE Transactions on Multimedia*, *10*(1), 86–96.
Zhu, X., Kandola, J., Ghahramani, Z., & Lafferty, J. (2005). Nonparametric transforms of graph kernels for semi-supervised learning. In *Advances in neural information processing systems* (pp. 1641–1648). Cambridge, MA: MIT Press.

Zenglin Xu is currently a doctoral student in the Department of Computer Science and Engineering in the Chinese University of Hong Kong, working with Professor Irwin King and Professor Michael R. Lyu. His research interests include machine learning and its applications to information retrieval, web search and data mining. He has published papers in several top conferences, such as NIPS, CIKM, ICDM, etc. Mr. Xu received his B.S. degree from Xi'an Polytechnic University in 2002 and his M.S. degree from Xi'an Jiaotong University in 2005, both in computer science and Engineering. During the summers of 2007 and 2008, he was a visiting student of Professor Rong Jin in Michigan State University, working on the problems of semi-supervised learning and kernel learning. He is a student member of IEEE Computer Society, ACM, and INNS.



Kaizhu Huang is currently a postdoctoral researcher at the University of Bristol, United Kingdom. He received the B.E. (1997) in Automation from Xi'an Jiaotong University, the M.E. (2000) in Pattern Recognition and Intelligent Systems from Institute of Automation, the Chinese Academy of Sciences, and the Ph.D. (2004) in Computer Science and Engineering from the Chinese University of Hong Kong. From 2004 to 2007, he was a researcher in the Information Technology Laboratory, Fujitsu Research and Development Center Co. Ltd. His research interests include machine learning, pattern recognition, image processing, and information retrieval.



Jianke Zhu received the Bachelor's degree in mechanical and electronics engineering from Beijing University of Chemical Technology, Beijing, PR China, in 2001, and the Master's degree in electrical and electronics engineering from the University of Macau, in 2005. He is currently a Ph.D. candidate in the department of computer science and engineering in the Chinese University of Hong Kong. His research interests are in computer vision, pattern recognition, image retrieval, and machine learning.

Irwin King received the B.Sc. degree in Engineering and Applied Science from the California Institute of Technology, Pasadena, in 1984. He received his M.Sc. and Ph.D. degree in Computer Science from the University of Southern California, Los Angeles, in 1988 and 1993 respectively. He joined the Chinese University of Hong Kong in 1993. His research interests include machine learning, multimedia processing, and web intelligence. In these research areas, he has published over 140 refereed journal and conference manuscripts. In addition, he has contributed over 20 book chapters and edited volumes. He is a senior member of IEEE and a member of ACM, International Neural Network Society (INNS), and Asian Pacific Neural Network Assembly (APNNA). Currently, he is serving the Neural Network Technical Committee (NNTC) and the Data Mining Technical Committee under the IEEE Computational Intelligence Society (formerly the IEEE Neural Network Society). He is also a governing board member of the APNNA. He is an Associate Editor of the IEEE Transactions on Neural Networks (TNN). He is a member of the Editorial Board of the Open Information Systems Journal, Journal of Nonlinear Analysis and Applied Mathematics, and Neural Information Processing–Letters and Reviews Journal (NIP-LR). He has also served as Special Issue Guest Editor for Neurocomputing and Journal of Computational Intelligent Research. He has served as program and/or organizing member in international conferences and workshops, e.g., WWW, ACM MM, ICME, ICASSP, IJCANN, ICONIP, ICPR, etc. He has also served as reviewer for international conferences as well as journals, e.g., Information Fusion, SIGMOD, IEEE TCAS, TNN, TPAMI, TMM, TKDE, TSMC, etc.

Michael R. Lyu received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981; the M.S. degree in computer engineering from University of California, Santa Barbara, in 1985; and the Ph.D. degree in computer science from University of California, Los Angeles, in 1988. He is currently a Professor in the Computer Science and Engineering department of the Chinese University of Hong Kong. He worked at the Jet Propulsion Laboratory as a Technical Staff Member from 1988 to 1990. From 1990 to 1992 he was with the Electrical and Computer Engineering Department at the University of Iowa as an Assistant Professor. From 1992 to 1995, he was a Member of the Technical Staff in the Applied Research Area of the Bell Communications Research, Bellcore. From 1995 to 1997 he was a research Member of the Technical Staff at Bell Laboratories, which was first part of AT&T, and later became part of Lucent Technologies. His research interests include software reliability engineering, distributed systems, fault-tolerant computing, web technologies, mobile networks, digital video library, multimedia processing, and video searching and delivery. Dr. Lyu has published over 300 refereed journal and conference papers in his research areas. He has been an associated editor of IEEE Transactions on Reliability, IEEE Transactions on Knowledge and Data Engineering, and Journal of Information Science and Engineering. He was elected to IEEE Fellow (2004) and AAAS Fellow (2007) for his contributions to software reliability engineering and software fault tolerance.