Nonlinear Dimensionality Reduction for Classification Using Kernel Weighted Subspace Method

Guang Dai, Dit-Yan Yeung
Department of Computer Science, Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
daiguang@cs.ust.hk, dyyeung@cs.ust.hk

Abstract--We study the use of kernel subspace methods that learn low-dimensional subspace representations for classification tasks. In particular, we propose a new method called *kernel weighted nonlinear discriminant analysis* (KWNDA) which possesses several appealing properties. First, like all kernel methods, it handles nonlinearity in a disciplined manner that is also computationally attractive. Second, by introducing weighting functions into the discriminant criterion, it outperforms existing kernel discriminant analysis methods in terms of the classification accuracy. Moreover, it also effectively deals with the small sample size problem. We empirically compare different subspace methods with respect to their classification performance of facial images based on the simple nearest neighbor rule. Experimental results show that KWNDA substantially outperforms competing linear as well as nonlinear subspace methods.

1. INTRODUCTION

Subspace methods play an important role in many pattern classification tasks. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two classical linear subspace methods that have been widely used for dimensionality reduction and feature extraction [1-7]. For solving classification problems, LDA-based algorithms generally outperform PCA-based ones since the former approach finds low-dimensional representations that maximize the differences between classes while the latter seeks low-dimensional representations that simply maximize data variance without taking into account class label information. However, many LDA-based algorithms suffer from the small sample size problem. The traditional solution to this problem is to utilize PCA concepts in conjunction with LDA, as was done for example in Fisherfaces [2]. Recently, more effective solutions, sometimes referred to as direct LDA (DLDA) methods, have been proposed [3-5]. While LDA-based methods have been demonstrated to perform well on many pattern classification applications, their performance is unsatisfactory when applied to problems that require nonlinear decision boundaries. For example, they are inadequate for many face recognition applications since high variability in facial features, including illumination, facial expressions and pose variations, incurs high nonlinearity in the representation space. Thus, it seems reasonable to assume that nonlinear generalizations of LDA should be able to provide better solutions.

The past decade has witnessed the emergence of some powerful methods called *kernel methods*, which make use of the so-called "kernel trick" to devise nonlinear generalizations of linear methods while preserving the computational tractability of their linear counterparts. The first such method is support vector machine (SVM), which essentially constructs a separating hyperplane in the high-dimensional (possibly infinite-dimensional) feature space F obtained through a nonlinear mapping $\phi : R^n \to F$. The kernel trick allows inner products in the feature space to be computed entirely in the input space without performing the mapping explicitly. Thus, for linear methods which can represent the relationships between data in terms of inner products only, they can readily be "kernelized" to give their nonlinear extensions [8-12]. As nonlinear extensions of PCA and LDA, kernel PCA (KPCA) and kernel nonlinear discriminant analysis (KNDA) have already been shown to provide better performance than their linear counterparts in several applications [8-12]. The basic ideas of KPCA and KNDA are to first map each input data point x into a feature space F via the nonlinear mapping ϕ and then perform PCA and LDA, respectively, in F. Similar to their linear counterparts, KNDA-based methods are generally better than KPCA-based methods for classification tasks. Mika et al. [9] first proposed a two-class kernel discriminant analysis algorithm, which was later generalized by Baudat and Anouar [10] to give the generalized discriminant analysis (GDA) algorithm for multi-class problems. However, same as traditional LDA-based algorithms, these KNDA-based algorithms also suffer from the small sample size problem and discard some significant discriminatory information. To overcome this problem, Lu et al. [11] proposed the kernel direct discriminant analysis (KDDA) algorithm. More recently, realizing the limitation of KDDA that some significant discriminatory information is still lost, we proposed a further enhanced algorithm called kernel generalized nonlinear discriminant analysis (KGNDA) algorithm [12].

However, for multi-class classification tasks, the existing discriminant analysis algorithms, including KGNDA, are not optimal as they tend to overemphasize the classes that are more separable. A solution to this problem is to introduce weighting functions to the discriminant criterion function [7], where a weighted between-class scatter matrix is defined to substitute for the conventional between-class scatter matrix. Classes that are closer to each other in the feature space, and thus can potentially impair the classification performance significantly, should be more heavily weighted in the input space. Based on this idea of incorporating weighting functions, we propose in this paper an improvement of KGNDA giving rise to a new method called kernel weighted nonlinear discriminant analysis (KWNDA). This new method possesses several appealing properties. First, like all kernel methods, it handles nonlinearity in a disciplined manner that is also computationally attractive. Second, by introducing weighting functions into the discriminant criterion, it outperforms existing kernel discriminant analysis methods in terms of the classification accuracy. Moreover, it also effectively deals with the small sample size problem. We apply KWNDA to face recognition where both the nonlinearity problem and the small sample size problem widely exist.

2. EXISTING KERNEL NONLINEAR DISCRIMINANT ANALYSIS METHODS

As discussed above, KNDA methods essentially perform LDA in the feature space F. Computation of the inner product of two vectors in F does not require applying the nonlinear

mapping ϕ explicitly when the kernel trick is applied through using an inner product kernel function:

$$k(x, y) = \phi(x)^T \phi(y) . \quad (1)$$

Let $x_i (i = 1, \ldots, N)$ denote one of N vectors in the training set X. X can be partitioned into c disjoint subsets X_i, i.e., $X = \bigcup_{i=1}^{c} X_i$, where X_i consists of N_i vectors that belong to class i with $N = \sum_{i=1}^{c} N_i$. The between-class scatter matrix S_b, within-class scatter matrix S_w, and population scatter matrix S_t in F can be expressed as follows:

$$S_b = \sum_{i=1}^{c} (N_i / N)(m_i - m)(m_i - m)^T \quad (2)$$

$$S_w = (1/N) \sum_{i=1}^{c} \sum_{x_j \in X_i} (\phi(x_j) - m_i)(\phi(x_j) - m_i)^T \quad (3)$$

$$S_t = S_b + S_w = (1/N) \sum_{i=1}^{N} (\phi(x_i) - m)(\phi(x_i) - m)^T \quad (4)$$

where $m_i = (1/N_i) \sum_{x_j \in X_i} \phi(x_j)$ denotes the sample mean of class i in F and $m = (1/N) \sum_{i=1}^{N} \phi(x_i)$ denotes the sample mean of all N vectors in F. We maximize the following criterion function to find the optimal coefficients w for the discriminants:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \quad (5)$$

KNDA algorithms that optimize (5) generally utilize the theory of reproducing kernels, which can express w as [8, 9]:

$$w = \sum_{i=1}^{N} \alpha_i \phi(x_i) \quad (6)$$

Substituting (6) into (5), $J(w)$ can be re-expressed as a function of $\alpha = (\beta_1, \ldots, \beta_N)$ as follows:

$$J_f(\alpha) = \frac{\alpha^T K_b \alpha}{\alpha^T K_w \alpha} \quad (7)$$

which is used as an equivalent criterion function for optimization, where K_b and K_w can be computed easily by applying the kernel trick. Details of the derivation can be found in [9].

Solving the above optimization problem is equivalent to solving a conventional eigenvalue problem to give the leading eigenvectors of $K_w^{-1} K_b$ with the largest eigenvalues. However, in the small sample case, K_w is not a full-rank matrix and hence is not invertible. A traditional solution used to circumvent this problem is to replace the inverse matrix K_w^{-1} with the pseudoinverse of K_w, as was done for example in GDA [10]. However, this method tends to lose the null space of K_w which potentially contains significant discriminatory information that can help to improve the classification accuracy.

To overcome the limitation of GDA, Lu et al. [11] proposed KDDA to apply the DLDA algorithm of Yu and Yang [3] in F. The underlying idea of KDDA is based on the assumption that discriminatory information in F only exists in the intersection space $(S_w(0) \bigcap S_b^{-1}(0))$, where $S_w(0) = \{x \mid S_w x = 0\}$ and $S_b^{-1}(0) = \{x \mid S_b x \neq 0\}$, and it intends to seek this intersection space. To do so, KDDA first computes $S_b^{-1}(0)$ by the eigenanalysis of S_b in F and then computes the discriminant coefficients by the eigenanalysis of the projection of S_w in $S_b^{-1}(0)$. Since the dimensionality of $S_b^{-1}(0)$ is no more than $c-1$, it is clear that the dimensionality of $(S_w(0) \bigcap S_b^{-1}(0))$ obtained in KDDA is less than $c-1$. However, according to [5, 12], it is obvious that the dimensionality of $(S_w(0) \bigcap S_b^{-1}(0))$ is $c-1$ in essence, since the dimensionality of F is far greater than the number of training examples in F. As a result, KDDA still discards some significant discriminatory information.

Recently, we proposed KGNDA to address the small sample size problem [12]. To prevent the loss of significant discriminatory information, the procedure of computing optimal discriminant coefficients in F, which essentially can be considered as a nonlinear extension of DLDA [4, 5], is carried out in KGNDA. KGNDA is based on the assumption that discriminatory information in F can be obtained from the intersection space $(S_w(0) \bigcap S_t^{-1}(0))$, since the intersection space $(S_w(0) \bigcap S_t^{-1}(0))$ is equivalent to the intersection space $(S_w(0) \bigcap S_b^{-1}(0))$ in practice, where $S_t^{-1}(0) = \{x \mid S_t x \neq 0\}$. To obtain $(S_w(0) \bigcap S_t^{-1}(0))$, KGNDA first computes $S_t^{-1}(0)$ by the eigenanalysis of S_t in F, and then obtains this intersection space by the eigenanalysis of the projection of S_w in $S_t^{-1}(0)$. Since $(S_w(0) \bigcap S_t^{-1}(0))$ can be obtained, KGNDA will compute the discriminant coefficients in this intersection space without losing the significant discriminatory information. More details can be found in [12].

3. KERNEL WEIGHTED NONLINEAR DISCRIMINANT ANALYSIS ALGORITHM

In this section, we present a new KWNDA algorithm to further improve the performance of KGNDA by incorporating weighting functions into the discriminant criterion of KGNDA.

The conventional discriminant criterion in (5) is not related to the classification ability of the feature representation obtained and overemphasizes the classes that are more separable in the input feature space F. As a result, the classification ability will be impaired. A solution to this problem is to introduce weighting functions to the discriminant criterion [7], where a weighted between-class scatter matrix is defined to replace the conventional between-class scatter matrix. Classes that are closer together in the output space, and thus can potentially impair the classification performance, should be more heavily weighted in the input space. According to [7], a weighted between-class scatter matrix S_B in F can be defined as:

$$S_B = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \frac{N_i N_j}{N^2} w(d_{i,j})(m_i - m_j)(m_i - m_j)^T \quad (8)$$

where $d_{i,j}$ is the Euclidean distance between the means of class i and class j in F, the weighting function $w(d_{i,j})$ is generally a monotonically decreasing function and $w(d_{i,j}) = d_{i,j}^{-p}$ with $p \in \mathbb{N}$. It is worthy to mention that matrix S_B equals matrix S_b under the special case of $w(d_{i,j}) = 1$. In addition, it is clear that $d_{i,j}$ in F can be calculated by the kernel trick as follows:

$$\begin{aligned} d_{i,j} &= \| m_i - m_j \| \\ &= \sqrt{\left(\sum_{i_1 \in X_i} \frac{\phi(x_{i_1})}{N_i} - \sum_{j_1 \in X_j} \frac{\phi(x_{j_1})}{N_j}\right)^T \left(\sum_{i_1 \in X_i} \frac{\phi(x_{i_1})}{N_i} - \sum_{j_1 \in X_j} \frac{\phi(x_{j_1})}{N_j}\right)} \\ &= \sqrt{\sum_{i_1, i_2 \in X_i} \frac{k_{i_1 i_2}}{N_i^2} + \sum_{j_1, j_2 \in X_j} \frac{k_{j_1 j_2}}{N_j^2} - \sum_{i_1 \in X_i, j_1 \in X_j} \frac{k_{i_1 j_1}}{N_i N_j} - \sum_{i_1 \in X_i, j_1 \in X_j} \frac{k_{j_1 i_1}}{N_i N_j}} \end{aligned} \quad (9)$$

where $k_{i,j} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

Hence, the weighted discriminant criterion in F is expressed as:

$$J_1(w) = \frac{w^T S_B w}{w^T S_w w} \quad \text{or} \quad J_2(w) = \frac{w^T S_B w}{w^T S_T w} \quad (10)$$

where $S_T = S_B + S_w$. According to [11], it is obvious that the former is equivalent to the latter. As in KGNDA, we assume that the significant discriminatory information with respect to $J_2(w)$ only exists in the intersection space $S_w(0) \bigcap S_T^{-1}(0)$, where $S_T^{-1}(0) = \{x \mid S_T x \neq 0\}$. However, it is intractable to compute this intersection space, due to two reasons. First, $S_w(0)$ is computationally intractable since the dimensionality of F may be arbitrarily large or even infinite. Second, it is intractable to compute $S_T^{-1}(0)$ by the eigenanalysis of S_T, since $S_T = S_B + S_w$.

Fortunately, we have proven the following lemma to overcome this problem. The detailed proof is omitted here due to space limitation.

Lemma 1. The space $S_t^{-1}(0) = \{x \mid S_t x \neq 0\}$ is equivalent to the space $S_T^{-1}(0) = \{x \mid S_T x \neq 0\}$, where S_t is the conventional population scatter matrix and $S_T = S_B + S_w$.

From this lemma, $S_w(0) \bigcap S_T^{-1}(0)$ can be obtained by calculating $S_w(0) \bigcap S_t^{-1}(0)$. From KGNDA, this intersection space can be calculated by the eigenanalysis of S_w and S_t in F , as follows:

Eigenanalysis of S_t in F : To obtain $S_t^{-1}(0)$, we need to compute the orthonormal bases of $S_t^{-1}(0)$. Then, S_t in (4) can be rewritten as:

$$S_t = \sum_{i=1}^{N} \overline{\phi}(x_i)\overline{\phi}(x_i)^T = \Phi_t \Phi_t^T \tag{11}$$

where $\overline{\phi}(x_i) = \sqrt{1/N}\,(\phi(x_i) - m)$, $\Phi_t = [\overline{\phi}(x_1), \ldots, \overline{\phi}(x_N)]$.

According to [11, 12], the orthonormal bases of $S_t^{-1}(0)$ can be obtained by computing the corresponding orthonormal eigenvectors of S_t with positive eigenvalues. Since the dimensionality of F , denoted as $N^{'}$, could be arbitrarily large or even infinite, it is intractable to directly compute those orthonormal eigenvectors of the $N^{'} \times N^{'}$ matrix S_t . Fortunately, those orthonormal eigenvectors can be indirectly derived from the eigenvectors of $\Phi_t^T \Phi_t$ (with size $N \times N$).

For all training examples $\{\phi(x_i)\}_{i=1}^{N}$ in F , we can define an $N \times N$ kernel matrix K as follows:

$$K = (k_{i,j})_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}} \tag{12}$$

where $k_{i,j} = \phi(x_i)^T \phi(x_j)$. Hence, by the kernel trick, $\Phi_t^T \Phi_t$ can be expressed as:

$$\Phi_t^T \Phi_t = \frac{1}{N}(K - \frac{1}{N}(K\ I_{N\times N} + I_{N\times N}\ K) + \frac{1}{N^2} I_{N\times N}\ K\ I_{N\times N}) \tag{13}$$

where $I_{N\times N}$ is the $N \times N$ matrix with all terms being one. Let λ_i and e_i ($i = 1, \ldots, m$) be the ith positive eigenvalue and the corresponding eigenvector of $\Phi_t^T \Phi_t$, respectively. According to [12], it is clear that $v_i = \Phi_t e_i \lambda_i^{-1/2}$ ($i = 1, \ldots, m$) constitute the orthonormal bases of $S_t^{-1}(0)$.

Eigenanalysis of S_w in F : Projecting S_w into the subspace spanned by $v_i = \Phi_t e_i \lambda_i^{-1/2}$ ($i = 1, \ldots, m$), it is clear that the projection $\overline{S}_w$ of S_w in this subspace can be expanded as:

$$\overline{S}_w = V^T S_w V = E^T \Xi^T \Xi E \tag{14}$$

where $V = [v_1, \ldots, v_m]$, $E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$,
$\Xi^T = K/N - \mathrm{I}_{N\times N}\ K/N^2 - K\ \mathrm{A}_{N\times N}/N + \mathrm{I}_{N\times N}\ K\ \mathrm{A}_{N\times N}/N^2$,
where $\mathrm{A}_{N\times N} = diag(A_1, \ldots, A_m)$, A_i is the $N_i \times N_i$ matrix with all terms being $1/N_i$.

Let $P = [\gamma_1, \ldots, \gamma_l]$ be the corresponding eigenvectors of the zero eigenvalues of $\overline{S}_w$ (generally, $l = c - 1$), so it is clear that $S_w(0) \bigcap S_T^{-1}(0)$ can be spanned by VP . Then, the optimal discriminant vectors with respect to $J_2(w)$ can be calculated in $S_w(0) \bigcap S_T^{-1}(0)$ without the loss of significant discriminatory information. From KGNDA, since the between-class distance is equal to zero in $S_w(0) \bigcap S_T^{-1}(0)$, the weighted discriminant criterion $J_2(w)$ in (10) can be replaced by $\hat{J}_2(w) = P^T V^T S_B V P$. By the kernel trick, it can be expanded as:

$$\begin{aligned} \hat{J}_2(w) &= P^T V^T S_B V P \\ &= P^T E^T \left(\sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \left((\sqrt{N_i N_j}/N^{3/2})\ w(d_{i,j})\ Z_{i,j}^T\ Z_{i,j} \right) \right) E P \end{aligned} \tag{15}$$

where $P = [\gamma_1, \ldots, \gamma_l]$, $V = [v_1, \ldots, v_m]$, $E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$, $Z_{i,j}^T = [KL_i + HKL_j - KL_j - HKL_i]$, H is the $N \times N$ matrix with all terms being $1/N$, L_i is the $N \times 1$ matrix, where the terms corresponding to class i are $1/N_i$ and the remaining terms are zero. It is clear that the matrix $P^T V^T S_B V P$ is a tractable $l \times l$ matrix. Let z_i be the ith eigenvector of $P^T V^T S_B V P$, where $i = 1, \ldots, l$, sorted in decreasing order of the corresponding eigenvalue $\lambda_i^{'}$. According to KGNDA, it is clear that $Y_i = VPz_i (i = 1, \ldots, l)$ constitute the optimal discriminant vectors with respect to the weighted discriminant criterion $J_2(w)$ in F . This gives the new KWNDA algorithm. For an input pattern x , its projection into the subspace spanned by $\Theta = [Y_1, \ldots, Y_l]$ can be calculated by $z = \Theta^T \phi(x)$, and this expression can be rewritten via the kernel trick as follows:

$$\begin{aligned} z = (Pz_1, \ldots, Pz_l)^T \bullet \\ \sqrt{\frac{1}{N}} E^T (k(x, x_1) - \frac{1}{N}\sum_{i=1}^{N} k(x, x_i), \ldots, k(x, x_N) - \frac{1}{N}\sum_{i=1}^{N} k(x, x_N))^T \end{aligned} \tag{16}$$

where $E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$, $P = [\gamma_1, \ldots, \gamma_l]$.

Thus, KWNDA can give a low-dimensional representation with enhanced discriminatory power. Moreover, this method also effectively addresses the nonlinearity problem and the small sample size problem.

4. EXPERIMENTAL RESULTS

To assess the performance of the KWNDA method proposed in this paper, we conduct some face recognition experiments to compare KWNDA with other subspace methods. Note that typical face recognition tasks suffer from the small sample size problem and require nonlinear methods, which are particularly suitable for demonstrating the strengths of KWNDA. Our experiments are performed on a database of 1045 images corresponding to 67 subjects by combining the following two image sources:

1) 47 subjects are selected from the FERET database, with each subject contributing 10 gray-scale images.
2) 20 subjects are selected from the UMIST database, with a total of 575 gray-scale images.

The gray-level and spatial resolutions of all images are 256 and 92×112, respectively. There exist large variations in illumination, facial expressions, and pose variations. As a result, the distribution of the face image patterns is highly nonlinear and complex.

The database is randomly partitioned into two disjoint sets for training and testing. Five images per subject are randomly chosen for training while the rest for testing. Thus, the training and test sets consist of 335 and 710 images, respectively. For each feature representation obtained by one subspace method, we use a simple nearest neighbor classifier with the Euclidean distance measure to assess the classification performance. Each experiment is repeated 10 times and the average classification rates are reported below.

The first experiment compares KWNDA with several linear subspace methods for face recognition, including eigenfaces [1], Fisherfaces [2], enhanced Fisher linear discriminant model (EFM) [6], Yu and Yang's DLDA [3], and Huang et al.'s DLDA [4]. In addition, for the sake of simplicity, KWNDA uses the RBF kernel function ($k(z_1, z_2) = \exp(\| z_1 - z_2 \|^2 / 10^9)$) and weighting function $w(d) = d^{-4}$. The experimental results shown in Fig.1 reveal that, as expected, KWNDA outperforms the linear subspace methods significantly.

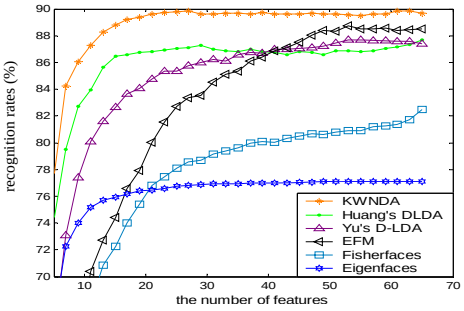

Fig.1. Comparative performance of KWNDA and several linear subspace methods, where the RBF kernel function $(k(z_1, z_2) = \exp(\| z_1 - z_2 \|^2 / 10^9))$ and weighting function $w(d) = d^{-4}$ are used for KWNDA.

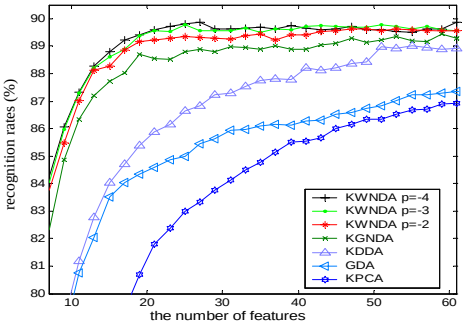

Fig.2. Comparative performance of KWNDA and several kernel nonlinear subspace methods, where the RBF kernel function $(k(z_1, z_2) = \exp(\| z_1 - z_2 \|^2 / 10^9))$ and different weighting functions $w(d) = \{d^{-2}, d^{-3}, d^{-4}\}$ are used for KWNDA.

Table 1 Average error percentages for different subspace methods when compared with KWNDA

Weighting function	d^{-2}	d^{-3}	d^{-4}
Huang's DLDA [4]	81.20%	79.91%	79.71%
EFM [6]	70.14%	69.11%	69.04%
KPCA [8]	70.14%	69.03%	68.88%
GDA [10]	75.65%	74.47%	74.32%
KDDA [11]	91.82%	90.40%	90.25%
KGNDA [12]	95.82%	94.31%	94.09%

The second experiment compares KWNDA with several kernel nonlinear subspace methods for face recognition, including KPCA [8], GDA [10], KDDA [11], and KGNDA [12]. Fig.2 shows the classification rates obtained for these different subspace methods. For our KWNDA method, the RBF kernel function $(k(z_1, z_2) = \exp(\| z_1 - z_2 \|^2 / 10^9))$ and different weighting functions $w(d) = \{d^{-2}, d^{-3}, d^{-4}\}$ are used. It can be seen that KGNDA is better than KPCA, GDA and KDDA, but KWNDA with different weighting functions further improve the performance of KGNDA. When the number of features is 30, the accuracy of KWNDA with weighting function $w(d) = d^{-4}$ is 89.6% while that of KGNDA is 88.1%. According to the definition of the average error percentage in [11], the average error percentage of the error rates of KWNDA over those of some other method is calculated as $\sum_{i=5}^{66}(1-\alpha_i)/(1-\beta_i)$, where α_i and β_i are the average recognition rates of KWNDA and one of the methods above, respectively, and i refers to the number of feature vectors. The average error percentages for different subspace methods are summarized in Table 1, showing that KWNDA is more effective than all other methods. More experiments have been performed by us but the results, which are consistent with those presented here, are not included in this paper due to space limitation. In general, all results obtained show than KWNDA outperforms the other subspace methods at least for the face recognition experiments we performed. The favorable results provide empirical justification for the discussions above which claim that KWNDA is able to generalize over and overcome some of the limitations of previous subspace methods.

5. CONCLUSION

We have presented a new kernel subspace method, called KWNDA, that learns low-dimensional representations for classification tasks. By incorporating weighting functions into the discriminant criterion, we improve upon the KGNDA method proposed by us before. Experimental results on face recognition show that KWNDA consistently outperforms existing linear and nonlinear subspace methods. Besides face recognition, we expect that KWNDA can also deliver excellent performance in other classification applications, including content-based image indexing and retrieval as well as video and audio classification.

6. ACKNOWLEDGMENTS

We would like to thank the authors of some previous papers for providing the code of their algorithms, including KPCA, GDA, and KDDA, to make direct comparison with our method possible. This research is supported by research grant CA03/04.EG01 (which is part of HKBU2/03/C) from the Research Grants Council of the Hong Kong Special Administrative Region, China.

7. REFERENCES

[1] M.Turk, A.P.Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol.3, no.1, pp.71–86, 1991.

[2] P.N.Belhumeur, J.P.Hespanha, D.J.Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, pp.711–720, July 1997.

[3] H.Yu, J.Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol.34, pp.2067–2070, 2001.

[4] R.Huang, Q.L.Liu, H.Q.Lu, S.D.Ma, "Solving the small sample size problem of LDA," *in Proceedings of the International Conference on Pattern Recognition*, pp.29-33, 2002.

[5] Y.F.Guo, L.D.Wu, "A novel optimal discriminant principle in high dimensional spaces," *in Proc. IEEE ICDL*, 2002.

[6] C.J.Liu, H.Wechsler, "Robust coding schemes for indexing and retrieval from large face recognition," *IEEE Transactions on Image Processing*, vol.9, pp.132-137, Jan, 2000.

[7] M.Loog, P.R.W.Duin, R.Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.23, no.7, pp.762-766, July, 2001.

[8] B.Scholkopf, A.Smola, K.R.Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol.10, pp.1299-1319, 1999.

[9] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, K.-R.Muller, "Fisher discriminant analysis with kernels," *in Proc. Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds:IEEE, 1999, pp.41–48.

[10] G.Baudat, F.Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol.12, pp.2385-2404, 2000.

[11] J.W.Lu, K.N.Plataniotis, A.N.Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Transactions on Neural Networks*, vol.14, pp.117-126. 2003.

[12] G.Dai, Y.T.Qian, "Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition," In *Proceedings of the IEEE International Conference on Image Processing*, pp.2697-2700. 2004.