FACE RECOGNITION USING FEATURE OPTIMIZATION AND ν-SUPPORT VECTOR LEARNING

Juwei Lu, K.N. Plataniotis, A.N. Venetsanopoulos
Multimedia Laboratory
Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto, Toronto, M5S 3G4, ONTARIO, CANADA

Abstract.

A new face recognition system is introduced and analyzed in this paper. The system utilizes a novel statistical pattern recognition method to optimize the feature selection process. The optimized feature set feeds a classification module, which is based on the modified ν-support vector machine approach (ν-SVM). The optimized feature set reduces the burden of the subsequent ν-SVM classifier and improves its learning speed and classification accuracy. The paper includes, simulation studies and comparative evaluation with several existing systems on the ORL face database. Results indicate that the proposed system has excellent performance achieving the lowest error rate reported to date for the ORL face database using only a very small set of features.

INTRODUCTION

Within the last decade, numerous algorithms have been proposed for face recognition. A number of surveys in this area can be found in [13, 17, 4]. Two issues are central to all these algorithms: (i) what features to use to represent a face, and (ii) how to classify a new face image based on the chosen feature representation.

The first issue addresses the problem of feature representation, where techniques available for its solution can be divided into two classes [9]: (i) geometric feature (GF) based methods [13], and (ii) statistical pattern recognition (SPR) based methods [17, 4]. GF-based methods [13] use properties and relations between facial characteristics such as eyes, nose, mouth and chin, as descriptors of faces for recognition. This class of methods relies heavily on the detection of facial features. Unfortunately, facial feature detection techniques developed to date have not been reliable enough to cater to this need [9]. In contrast with GF-based methods, SPR-based methods generally operate directly on the pixel level and process face images as holistic patterns. Since there is no need to detect facial features, this class of

approaches is demonstrated to be more robust and practical comparing to GF-based methods [9, 17]. The most successful SPR-based methods include Eigenfaces [16] and Fisherfaces [1].

Although research studies indicate that linear discriminant analysis (LDA) based algorithms [1, 5, 20] outperform principal component analysis (PCA) based algorithms [16] in many applications of practical importance such as face recognition, traditional LDA algorithms cannot provide reliable and robust solutions since their separability criteria are not directly related to their classification accuracy in the output space, often resulting in misclassification [11]. As a result, weighting functions are introduced in LDA based techniques [11]. Object classes that are closer together, and thus can potentially result in misclassification, are more heavily weighted. Based on the idea, we utilize a new LDA algorithm, called DF-LDA [12], for feature extraction in this paper. DF-LDA optimizes the feature space utilizing a fractional-step dimensionality reduction algorithm which allows for the relevant distances in LDA to be more accurately weighted. DF-LDA also provides more reliable and robust solutions to the so-called "small sample size problem" (SSS) which exists in high-dimensional pattern recognition tasks where the number of samples is smaller than the dimensionality of the samples [1, 5, 20]. DF-LDA can been seen as an integrated LDA approach, which takes advantages of the existing LDA techniques while at the same time overcomes many of their shortcomings and limitations.

The another issue addresses the problem of pattern classification. Taking the optimized feature vectors as the input, a modified ν-support vector machine (ν-SVM) [3] is trained to learn non-linear boundaries separating different classes of patterns. Support vector machines (SVM) comprise a new class of learning algorithms, originally developed by Vapnik [18]. SVMs differ radically from comparable approaches such as neural networks, since they embody the Structural Risk Minimization principle, which has been shown to be superior to traditional Empirical Risk Minimization principle employed by most of neural networks in many classification and regression problems [2, 7]. In conventional SVMs, the only parameter that we can dispose of is the regularization constant C, which has important effect on the performance of SVMs. However, it is difficult to select an appropriate C. Schölkopf [14] presented a new support vector algorithm called "ν-SVM", where C is substituted by a parameter ν. ν lets one be able to effectively control the number of support vectors and errors. However, comparing to regular SVMs, ν-SVM is more complicated so that there are no effective methods to solve large-scale ν-SVM. In this paper, we adopt a variant of ν-SVM proposed by Chang [3] in which ν-SVM is modified to a different form where existing methods can be applied to solve it, while advantages of regular SVMs are preserved.

The organization of the rest of the paper is as follows. In Section 2, we introduce the DF-LDA algorithm and the ν-SVM based classifier. In Section 3, two sets of experiments are presented to demonstrate advantages of the proposed system. In which, DF-LDA is compared with Eigenfaces, Fisherfaces and direct LDA algorithm (D-LDA) [20] based on a common

classifier. Then, ν-SVM is also compared with other 3 classifiers: nearest neighbor (NN), nearest feature line (NFL) [9] and probabilistic neural networks (PNN) [19] taking the same feature set as input. The results show that both of DF-LDA and ν-SVM obtain the best performance in terms of error rate. Moreover, our system consisting of DF-LDA and ν-SVM achieves the lowest error rate reported to date for the ORL face database using only a very small set of features. Finally, Section 4 gives the conclusions.

METHODS

Given a set of training face images $\{\mathbf{z}_i\}_{i=1}^{L}$, each of images is represented as a vector of length $N(= I_w \times I_h)$, $i.e.$ $\mathbf{z}_i \in \mathbb{R}^N$, and belongs to one of c classes $\{\mathbf{Z}_i\}_{i=1}^{c}$, where L is the number of face images, $I_w \times I_h$ is the face image size, and $\mathbb{R}^N$ denotes N-dimensional real space. The processing on these patterns consists of two parts: the SPR-based optimal feature extraction and the ν-SVMs based learning for classification.

SPR-based Feature Selection

Traditional LDA-based algorithms. LDA is one of the most popular SPR-based techniques for feature extraction. Researchers have demonstrated that the LDA based algorithms outperform the PCA based algorithm for solving problems of pattern classification [1, 5, 20]. Because LDA optimizes the low-dimensional representation of the objects with focus on the most discriminant feature extraction while PCA achieves simply object reconstruction.

Let S_{BTW} and S_{WTH} be the between- and within-class scatter matrices respectively, LDA found the most discriminant basis vectors, denoted by $\{\psi_k\}_{k=1}^{M}$, by $\Psi = \arg\max_{\Psi} \frac{|(\Psi^T S_{BTW} \Psi)|}{|(\Psi^T S_{WTH} \Psi)|}$ where $\Psi = [\psi_1, \ldots, \psi_M]$ and $M < N$. Assuming that S_{WTH} is non-singular, the basis vectors $\{\phi_k\}_{k=1}^{M}$ correspond to the first M eigenvectors with the largest eigenvalues of $S_{WTH}^{-1} S_{BTW}$.

However, in face recognition tasks the number of training samples L is, in most cases, much smaller than the dimensionality N leading to a degenerated S_{WTH}, which produced the SSS problem. Many methods have been proposed to solve the problem [1, 5, 20]. A typical solution adopted in Fisherfaces [1] is, to perform an intermediate dimensionality reduction before doing the FLD analysis by PCA transformation to remove null space from S_{WTH}. Nevertheless, the Fisherface method inevitably suffer the same drawback as the Eigenface method in the PCA step, which leads some of significant discriminatory information may be lost [20].

Aimed to the problem with Fisherface method, Yang et al. [20] and Chen et al. [5] proposed direct LDA algorithms (D-LDA) respectively without a separate PCA step. The basic idea of D-LDA is that the null space of S_{WTH} may contains significant discriminatory information if the projection of S_{BTW} is not zero in that direction, but it will not lose any useful information if the

null space of S_{BTW} is discarded. However, a drawback with the D-LDA method is that zero eigenvalues of S_{WTH} may be used as divisors [12].

Integrated LDA (DF-LDA). A disadvantage of traditional LDA algorithms including Fisherfaces and D-LDA is that the separability criteria does not have a direct relationship with the classification accuracy in the output space [11]. The weighting function on the between-class scatter helps make the optimality criteria more representative of the classifiability in the output space based on the motivation that classes which are closer to one another are more likely to cause confusion and thus should be more heavily weighted. Let $\hat{S}_{BTW}$ be the weighted between-class scatter matrix in the input space and given by, $\hat{S}_{BTW} = \sum_{i=1}^{c} \phi_i \phi_i^T$ where, $\phi_i = \sum_{j=1}^{c} (L_i \cdot w(d_{ij}))^{1/2} (\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j)$ and $d_{ij} = \parallel \bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j \parallel$ is the Euclidean distance between the means of class i and class j in the input space. The weighting function $w(d_{ij})$ should be a monotonically decreasing function because classes that are closer together should be given a greater weightage.

During dimensionality reduction using LDA, another possible problem will be encountered. Suppose we want to reduce the dimensionality of face images $\{\mathbf{z}_i\}_{i=1}^{L}$ from N to $(N-1)$. We firstly diagonalize $\hat{S}_{BTW}$, and get its eigenvectors, $[v_1, \ldots, v_N]$. The required lower dimensional representation of $\mathbf{z}_i$ is obtained by projecting $\mathbf{z}_i$ into the subspace spanned by the first $(N-1)$ most significant eigenvectors $[v_1, \ldots, v_{N-1}]$. However, it is possible that there is a pair of classes i and j whose difference $(\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j)$ has approximately the same orientation as the discarded eigenvector v_N. Thus, the two classes would heavily overlap in the $(N-1)$-dimensional subspace, which is orthogonal to v_N. The problem is caused by insufficiently weighting on $(\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j)$ when calculating $\hat{S}_{BTW}$ [11], *i.e.* classes i and j are well-separated in the input space and this leads a small $w(d_{ij})$. Instead of reducing the dimensionality from N to $N-1$ directly, DF-LDA uses a fraction-step LDA (F-LDA) like algorithm which move "incrementally" towards a dimensionality of $(N-1)$. At each incremental step, $\hat{S}_{BTW}$ is recomputed and rediagonalized, thereby allowing those classes that come closer together to be increasingly weighted and causing the $(N-1)$-dimensional subspace to reorient and avoid severe overlap between classes in the output space.

However, the SSS problem with face recognition tasks leads to degenerated S_{WTH} and $\hat{S}_{BTW}$. It is not applicable to directly apply F-LDA to find $\{\psi_k\}_{k=1}^{M}$. Therefore, DF-LDA implements feature optimization in two steps. Firstly, an improved D-LDA algorithm is used to remove the null space of $\hat{S}_{BTW}$ and solve SSS problem without losing the most discriminant features. This also lower dimensionality of feature space from N to $M' < N$. Then, feature space is further optimized using a F-LDA like technique. Finally, a lower-dimentional feature space spanned by $M < M'$ most discriminant feature basis vectors can be obtained. DF-LDA algorithm naturally unifies the D-LDA and F-LDA techniques together, and takes advantages of the two techniques while at the same time overcomes many of their shortcomings and limitations. The detail pseudocode to implement the DF-LDA algorithm can

be found in [12]. The first 6 of M when $M = 25$ and $M = 10$ in our experiments are shown in Fig.1 (Right) respectively. Let Γ denote the projection matrix consisting of the M basis vectors obtained by DF-LDA. For each face image $\mathbf{z}_i$, a feature vector $\mathbf{x}_i$ can be calculated by $\mathbf{x}_i = \Gamma^T \mathbf{z}_i$, which is then fed into subsequent classifiers.
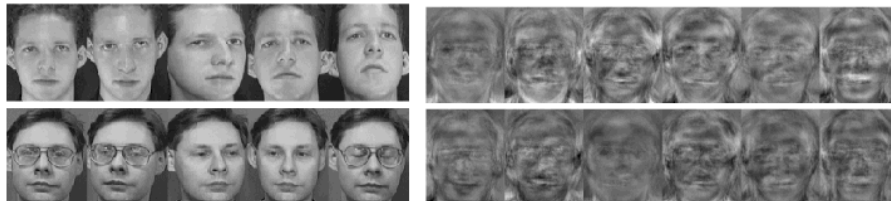

Figure 1: **Left:** Sample images of two persons in the ORL database. **Right:** The optimal basises derived by the DF-LDA method, top: the first 6 basises obtained when $M = 25$, $r = 30$ and $w(d) = d^{-8}$, bottom: the first 6 basises obtained when $M = 10$, $r = 30$ and $w(d) = d^{-8}$, both of them are listed from left to right.

ν-SVM based classifier

Most methods for training a classifier, such as neural networks and radial basis function (RBF), are based on the principle of minimizing Empirical Risk, *i.e.* training error. As a result, the classifiers obtained may be entirely unsuitable for classification of unseen test patterns, although they may achieve lowest training error. In contrast, SVMs embody another induction principle, *i.e.* the Structural Risk Minimization, which aims to minimize an upper bound on the expected generalization error. Since the notation and terminology of SVM approaches still remain relatively unknown, a brief overview on conventional C-SVM is included here.

C-**Support vector classification.** Given training vectors $x_i \in \mathbb{R}^n$, $i = 1, \ldots, l$ belonging two classes, and a class label $y_i \in \{1, -1\}$ for every x_i, the general optimal separating hyperplane given in C-SVM is the one that,

$$Minimize: \quad P_C = \frac{1}{2} \left\| \mathbf{w} \right\|^2 + C \sum\nolimits_{i=1}^{l} \xi_i \tag{1}$$

subject to $y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i$, $\quad \xi_i \geq 0, i = 1, \ldots, l$, where ξ_i are slack variables, C is a regularization constant and the hyperplane of equation is defined by parameters $\mathbf{w}$ and b.

 Our target is to find the solution of $(\mathbf{w}, b)$ from Equ(1) and its constraints. The solution can be given by the saddle point of the Lagrange function,

$$L_C = \frac{1}{2} \left\| \mathbf{w} \right\|^2 + \sum\nolimits_{i=1}^{l} \{ C\xi_i - \alpha_i(y_i(\mathbf{w}^T x_i + b) - 1 + \xi_i) - \mu_i \xi_i \} \tag{2}$$

where α_i and μ_i are positive Lagrange multipliers. Then the solution can be found by minimizing L_C with respect to $\mathbf{w}$, b, ξ_i and simultaneously

requiring $\frac{\partial L_C}{\partial \alpha_i} = 0$, $\frac{\partial L_C}{\partial \mu_i} = 0$, all subject to $\alpha_i \geq 0$ and $\mu_i \geq 0$. According to the Wolf dual [6], the above *primal* Lagrangian problem can be transformed to its *dual* problem, which becomes:

$$Maximize: \quad D_C = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (3)$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$.

We can easily obtain the solution of the primal problem from the dual, and denote them as $(\mathbf{w}^*, b^*)$, where $\mathbf{w}^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i$.

For a new data point x, the classification is given by a decision function,

$$f_C(x) = sign(\sum_{i=1}^{l} (\mathbf{w}^* \cdot x_i + b^*)) = sign(\sum_{i=1}^{l} (y_i \alpha_i^* x \cdot x_i + b^*)) \quad (4)$$

In the case where the decision function is not a linear function of the data, SVMs firstly map the input vector, x, into a high dimensional feature space by a nonlinear function $\phi(x)$, and then construct an optimal separating hyperplane in the higher dimensional space (possibly infinite dimensional). Fortunately, the exact $\phi(x)$ is not needed. The only requirement is to replace the inner product $< x_i \cdot x_j >$ in the input space with a kernel function $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ [18]. Since the dual (Equ(3)) and the decision function (Equ(4)) only depend on the input data x through dot products. There are three typical kernel functions: Polynomial, Gaussian Radial Basis Function and Multi-Layer Perception [18].

ν-Support vector classification. The original ν-SVM [14] is more complicated than regular C-SVM so that there are no effective methods to solve large-scale ν-SVM [3]. Therefore, we adopt a variant of ν-SVM proposed by Chang [3] in which ν-SVM is modified to a different form where existing methods can be applied to solve it, while advantages of the original ν-SVM are preserved.

The primal form in the modified ν-SVM [3] is:

$$Minimize: \quad P_\nu = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{2} b^2 - \nu\rho + \frac{1}{l} \sum_{i=1}^{l} \xi_i \quad (5)$$

subject to $y_i(\mathbf{w}^T \phi(x_i) + b) \geq \rho - \xi_i$, $\xi_i \geq 0, i = 1, \ldots, l$, $\rho \geq 0$.

The dual of P_ν is:

$$Maximize: \quad D_\nu = -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (K(x_i, x_j) + 1) \quad (6)$$

subject to $0 \leq \alpha_i \leq 1/l$, $\sum_{i=1}^{l} \alpha_i = \nu$.

Compared with the original ν-SVM [14], constraints of the dual (D_ν) here is simplified so that D_ν has a similar form to D_C and can be solved by the same way as D_C.

The decision function is then given by,

$$f_\nu(x) = sign(\sum_{i=1}^{l} y_i \alpha_i^* (K(x, x_i) + 1)) \quad (7)$$

In the original SVM algorithm [18], the only parameter that we can dispose of is the regularization constant C (that is why it is called C-SVM). C has important effects on the performance of SVMs. Unfortunately, it is difficult to select an appropriate C in practical application, since we can not give an intuitive interpretation on C. C is substituted by parameter ν in ν-SVM, which is proved to be an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors in [14]. An advantage of ν-SVM [14] over C-SVM is that ν lets one be able to effectively control the number of support vectors and errors instead of C.

Although the ν-SVM classifiers described above are binary classifiers, they are easily combined to handle the multiclass case. A simple and effective method is to train c one-versus-rest classifiers for the c-class case. Let $(\alpha_{ij}^*, x_{ij}^*, y_{ij}^*)$ denote boundary parameters of j-th classifier, where $j = 1 \cdots c$, $i = 1 \cdots m_j$, $m_j \leq l$ is number of support vectors, and (x_{ij}^*, y_{ij}^*) are $i - th$ support vector and its class label. For a test point x, its distance from the boundary is calculated by

$$d_j(x) = \sum\nolimits_{i=1}^{m_j} y_{ij}^* \alpha_{ij}^* (K(x, x_{ij}^*) + 1) \tag{8}$$

The decision is then given by $k^* = \arg\max_j d_j(x)$, and k^* is the class for x to be that corresponding to the largest distance.

EXPERIMENTAL RESULTS

Two sets of experiments are presented to demonstrate the performance of our system. One is to compare DF-LDA with several existing feature selection algorithms: Eigenfaces [16], Fisherfaces [1] and D-LDA algorithm [20] in terms of the error rate based on a simple nearest neighbor classifier (NN). Another is to compare the ν-SVM with other 3 classifiers: NN, nearest feature line (NFL) [9], and probabilistic neural networks (PNN) [19] based on the same input feature set. Both of them are conducted on ORL Cambridge face database [15].

The ORL database contains 40 distinct persons, each person having 10 different images, taken at different times, varying lighting slightly, facial expressions and facial details. Fig.1 (Left) shows sample images of two randomly selected persons in the database.

Face recognition procedure is performed for the two experiments in the following two stages: (1) Feature extraction. The training set and query set are derived in the same way as in [9, 8]: The 10 images of each of the 40 persons is randomly partitioned into two sets, resulting in 200 training images and 200 testing images, with no overlapping between the two. Then, both of them are projected into the feature space derived from Eigenface, Fisherface, D-LDA and DF-LDA methods. (2) Classification. Taking feature vectors derived from (1) as feeds, classifiers are trained and tested by training and testing sets of feature vectors respectively. The final error rates are the

averages of the error rates obtained by 8 runs (4 runs in [9] and 3 runs in [8]), each run being performed on a random partition of the database into two sets.

DF-LDA vs Eigenfaces, Fisherfaces and D-LDA

This experiment compares the DF-LDA with several popular feature selection algorithms used for face recognition: Eigenfaces [16], Fisherfaces [1] and D-LDA [20] based on a simple NN classifier. $r = 30$ and $w(d) = d^{-8}$ is used as the number of fractional steps and the weighting function respectively, and dimensionality of face images is reduced from $N = 10304$ to $M = 25$ in the DF-LDA method.
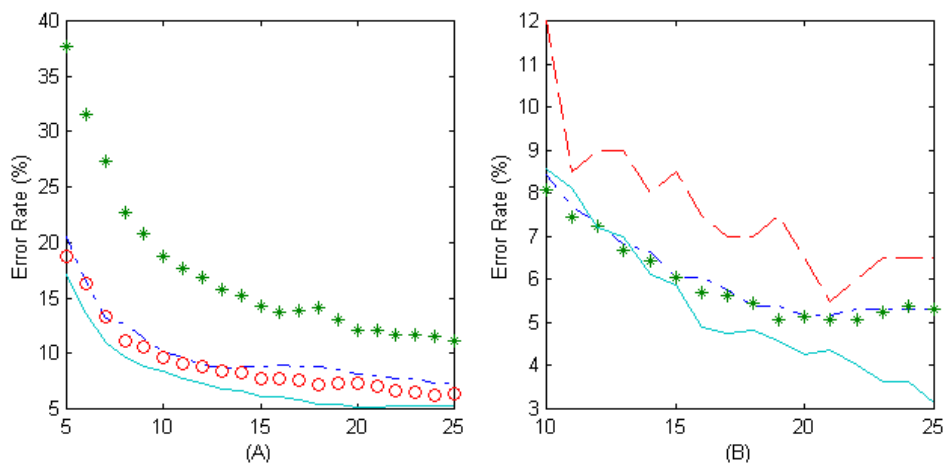

Figure 2: Comparison of error rates as functions of the number of feature vectors. **A:** '-.': Eigenfaces, '*': Fisherfaces, 'o': D-LDA and '-': DF-LDA using NN based classifiers. **B:** '-.': NN, '*': PNN, '- -': NFL and '-': ν-SVM

The error rate as functions of the number of feature vectors obtained by the four methods are given in Fig.2 (A), in which the performance of DF-LDA is overall superior to other three methods. Let α_i and β_i be the error rate of DF-LDA and one of other three methods respectively, where $i = [5 \ldots 25]$ is the number of feature vectors. We can obtain the average percentage of the error rate of DF-LDA over that of other methods by $\sum_{i=5}^{25} (\alpha_i/\beta_i)$. The results show that the average error rate of the DF-LDA is only about 73.32%, 43.36% and 80.77% of that of the Eigenfaces, Fishfaces and D-LDA algorithms respectively. Also, of interest is the performance of Eigenfaces vs Fishfaces. Not surprisingly, Eigenfaces outperform Fisherfaces, because Fishfaces may lost significant discriminant information due to the intermediate PCA step. The similar observation can be found in [10].

ν-SVM vs NN, NFL and PNN

Taking the feature vectors derived from the DF-LDA algorithm as input, this experiment compares performance among different classifiers: ν-SVM, NN, NFL [9] and PNN [19]. The error rate as functions of the number of feature vectors obtained by the four methods are given in Fig.2 (B), in which ν-SVM obtains the best performance when RBF kernel ($K(x,y) = exp(-\beta \|x - y\|^2)$), $\beta = 2.9$ and $\nu = 0.09$ are used. The average error rate of the ν-SVM calculated as the same way as last experiment, is only about 85.74%, 87.74% and 69% of that of the NN, PNN and NFL classifiers respectively. Especially when 25-dimensional DF-LDA feature vectors are used, the error rate of ν-SVM is only 3.125%, and it achieves the lowest error rate reported previously by [9, 8] with the least dimensional feature vectors used (40-dimensional feature vectors (based on Eigenfaces) are used in [9]).

CONCLUSION

A new face recognition system based on the two new techniques: DF-LDA and ν-SVM has been proposed. The feature set is optimized for the best linear discrimination by DF-LDA. Although the optimization may not be optimal for nonlinear discrimination as performed by the subsequent ν-SVMs, it provides more compact clusters in the optimized feature space and better separability between clusters. We believe that this would help ease the training of the ν-SVMs and obtain better classification performance.

ν-SVMs are used to learn a complex function for adequate classification given the feature set. This is advantageous over the SPR approach such as NN, NFL and PNN in that a SPR-based classifier takes a functional form which is often too simplified for mathematical tractability at sacrifice of classification performance, and has to tackle parameter estimation which is often performed on *ad hoc* basis. Results indicate that the proposed system has excellent performance and achieves the lowest error rate reported to date for the ORL face database using only a very small set of features.

The proposed system has provided a general platform to solve high-dimensional pattern recognition problems. We expect that it will have also good performance in other applications such as content-based image indexing, video and audio classification.

REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, ""Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection"," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 19, no. 7, pp. 711–720, July 1997.

[2] C. J. C. Burges, ""A Tutorial on Support Vector Machines for Pattern Recognition"," **Knowledge Discovery and Data Mining**, vol. 2, 1998.

[3] C.-C. Chang and C.-J. Lin, ""Training nu-Support Vector Classifiers: Theory and Algorithms"," **Neural Computation**, 2001.

[4] R. Chellappa, C. Wilson and S. Sirohey, ""Human and machine recognition of faces: A survey"," **Proceedings of the IEEE**, vol. 83, pp. 705–740, 1995.

[5] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin and G.-J. Yu, ""A new LDA-based face recognition system which can solve the small sample size problem"," **Pattern Recognition**, vol. 33, pp. 1713–1726, 2000.

[6] R. Fletcher, **Pratical Methods of Optimization**, John Wiley and Sons, Inc., 2nd edn., 1987.

[7] K. Jonsson, J. Matas, J. Kittler and Y. Li, ""Learning Support Vectors for Face Verification and Recognition"," in **Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000**, Grenoble, France, March 2000.

[8] S. Lawrence, C. L. Giles, A. Tsoi and A. Back, ""Face Recognition: A Convolutional Neural Network Approach"," **IEEE Transactions on Neural Networks**, vol. 8, no. 1, pp. 98–113, 1997.

[9] S. Z. Li and J. Lu, ""Face Recognition Using the Nearest Feature Line Method"," **IEEE Transactions on Neural Networks**, vol. 10, pp. 439–443, March 1999.

[10] C. Liu and H. Wechsler, ""Evolutionary Pursuit and Its Application to Face Recognition"," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 22, no. 6, pp. 570–582, June 2000.

[11] R. Lotlikar and R. Kothari, ""Fractional-Step Dimensionality Reduction"," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 22, no. 6, pp. 623–627, June 2000.

[12] J. Lu, K. Plataniotis and A. Venetsanopoulos, ""Face Recognition Using LDA Based Algorithms"," **submitted to IEEE Transactions on Neural Networks**, 2001.

[13] A. Samal and P. A.Iyengar, ""Automatic Recognition And Analysis of Human Faces and Facial Expressions: A Survey"," **Pattern Recognition**, vol. 25, pp. 65–77, 1992.

[14] B. Schölkopf, A. Smola, R. Williamson and P. Bartlett, ""New support vector algorithms"," **Neural Computation**, vol. 12, pp. 1083–1121, 2000.

[15] W. site of ORL Cambridge face database:, ""http://www.cam-orl.co.uk/facedatabase.html"," **AT&T Laboratories Cambridge**, 1994.

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

[17] D. Valentin, J. O. T. Herve Abdi Alice and G. W. Cottrell, ""Connectionist Models of Face Processing: A Survey"," **Pattern Recognition**, vol. 27, no. 9, pp. 1209–1230, 1994.

[18] V. Vapnik, **The Nature of Statistical Learning Theory**, Springer-Verlag, 1995.

[19] P. Wasserman, **Advanced Methods in Neural Computing**, New York: Van Nostrand Reinhold: Springer-Verlag, 1993.

[20] J. Yang, H. Yu and W. Kunz, ""An Efficient LDA Algorithm for Face Recognition"," in **Proceedings of Sixth International Conference on Control, Automation, Robotics and Vision**, Singapore, December 2000.