Journal of Korea Multimedia Society Vol. 13, No. 6, JUNE 2010(pp. 817-824)

An Ensemble Classifier using Two Dimensional LDA

Cheong Hee Park†

ABSTRACT

Linear Discriminant Analysis (LDA) has been successfully applied for dimension reduction in face recognition. However, LDA requires the transformation of a face image to a one-dimensional vector and this process can cause the correlation information among neighboring pixels to be disregarded. On the other hand, 2D-LDA uses 2D images directly without a transformation process and it has been shown to be superior to the traditional LDA. Nevertheless, there are some problems in 2D-LDA. First, it is difficult to determine the optimal number of feature vectors in a reduced dimensional space. Second, the size of rectangular windows used in 2D-LDA makes strong impacts on classification accuracies but there is no reliable way to determine an optimal window size. In this paper, we propose a new algorithm to overcome those problems in 2D-LDA. We adopt an ensemble approach which combines several classifiers obtained by utilizing various window sizes. And a practical method to determine the number of feature vectors is also presented. Experimental results demonstrate that the proposed method can overcome the difficulties with choosing an optimal window size and the number of feature vectors.

Key words: Dimension reduction, Ensemble classifier, Face recognition, Two-dimensional LDA

1. INTRODUCTION

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been successfully applied for dimension reduction in various application areas, especially for high dimensional data such as face recognition and text categorization [1,2]. Traditional dimension reduction methods are based on vector space representation of data. In face recognition, in order to apply PCA or LDA, one needs to transform a face image to a one-dimensional vector. The transformation is performed by stacking pixel intensity values column by column or row by row [3,4]. Based on the transformed one-dimensional vectors, optimal projective directions are searched and a reduced dimensional space is obtained by projecting the original data vectors onto the projective directions.

Since a transformation process rearranges the pixels of a 2D-image by rows, it may cause disregard of useful information such as the correlations among neighboring pixels. In recent years several algorithms which directly use 2D images without a transformation process have been proposed. This approach is called 2D-PCA or 2D-LDA [5,6]. Figure 1 describes the difference between traditional 1D-dimension reduction and 2D-dimension reduction. In the figures of the top line, W represents the projective direction vector obtained by 1D-dimension reduction. The coefficients of W were rearranged in the form of a matrix so that each coefficient is mapped to the corresponding pixels of a face image. In 2D-dimension reduction which is shown in the bottom line of Figure 1, a face image is partitioned to blocks. 2D-dimension reduction methods compute

※ Corresponding Author : Cheong Hee Park, Address : (305-764) 220, Gung-dong, Yuseong-gu, Daejeon, Korea
TEL : +82-42-821-6293, FAX : +82-42-822-4997, E-mail: cheonghee@cnu.ac.kr
Receipt date : May 26, 2009, Revision date : Nov. 10, 2009
Approval date : Mar. 31, 2010
† Chungnam National University Dept. of Computer Science and Engineering
※ This study was financially supported by research fund of Chungnam National University in 2008.

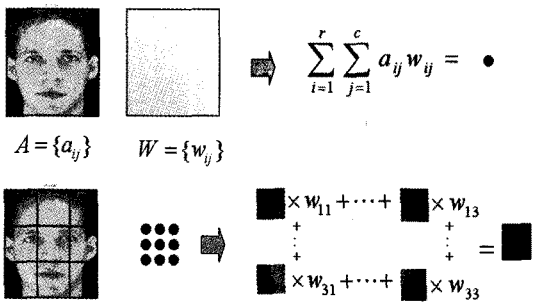


Fig. 1. Comparison of 1D-dimension reduction (top) and 2D-dimesion reduction (bottom).

a projective vector whose coefficients correspond to each block.

Although it has been shown that 2D-dimension reduction methods are superior to traditional PCA or LDA, there are some problems. First, unlike in LDA it is difficult to determine the optimal number of feature vectors extracted. Second, the size of rectangular windows used in 2D dimension reduction makes strong impacts on classification accuracies but there is no efficient method to choose a right window size. Cross-validation is usually adopted for parameter selection. However, cross-validation is expensive, since dimension reduction processes should be repeated for each hold-out case. It is also unreliable in undersampled problems where the number of samples is much smaller than a data dimension. In this paper, a practical method to determine the number of projective vectors is presented. It simplifies a leave-one-out procedure, resulting in the saving of computational complexities. We also propose an ensemble approach for 2D-LDA which utilizes various window sizes without selection processes.

The rest of the paper is organized as follows. In Section 2, a brief review of 2D-LDA is given. In Section 3, a practical method to determine the number of projective vectors and an ensemble approach for 2D-LDA are presented. Experimental results in Section 4 demonstrate that the proposed method overcomes the difficulties with choosing an optimal window size and achieves competent performances.

2. TWO-DIMENSIONAL LDA

In face recognition, dimension reduction has been proved to be very effective in dealing with high dimensionality of face images. Recently, two-dimensional PCA (2D-PCA) and two-dimensional LDA (2D-LDA) have been proposed by utilizing two dimensional subimage variances instead of pixel level variances [5,6].

Let us denote a given data set of face images as

$$[A_1^1, \cdots, A_{n_1}^1, \cdots\cdots, A_1^h, \cdots, A_{n_h}^h], \tag{1}$$

where A_i^j is a two-dimensional face image of a size $r \times c$ with a width of r and a height of c. When the data set consists of the face images of h subjects, A_i^j denotes the i-th image sample of the j-th subject and $n = \sum_{i=1}^{h} n_i$.

In two-dimensional dimension reduction, a face image is partitioned to blocks of a size $s \times t$. Let B_i^j denote the reshaped image of A_i^j whose columns are constructed by one dimensional representation of each block of A_i^j. Hence a height of B_i^j is $s \times t$ and a width is the number of blocks, $\frac{r}{s} \times \frac{c}{t}$. Now dimension reduction is performed by

$$b_i^j = B_i^j w, \tag{2}$$

which gives a column vector of a size $s \times t$. The righthand side of Eq. (2) can be interpreted as a linear combination of block images represented as the columns of B_i^j, weighted by the components in the projection vector w. When a block size $s \times t$ is 1×1, 2D-dimension reduction becomes equal to traditional 1D-dimension reduction.

In 2D-PCA, an optimal projective direction is the one to maximize a variance in the transformed space. By denoting $\overline{B} \equiv \frac{1}{n}\sum_{j=1}^{h}\sum_{i=1}^{n_j} B_i^j$ and $\bar{b} \equiv \frac{1}{n}\sum_{j=1}^{h}\sum_{i=1}^{n_j} b_i^j = \overline{B}w$, the variance is formulated as

$$trace\left(\sum_{j=1}^{h}\sum_{i=1}^{n_j} (b_i^j - \bar{b})(b_i^j - \bar{b})^T\right)$$

$$= trace\Big(\sum_{j=1}^{h}\sum_{i=1}^{n_j}(B_i^j w-\bar{B}w)(B_i^j w-\bar{B}w)^T\Big)$$

$$= \sum_{j=1}^{h}\sum_{i=1}^{n_j} trace\big((B_i^j w-\bar{B}w)(B_i^j w-\bar{B}w)^T\big)$$

$$= \sum_{j=1}^{h}\sum_{i=1}^{n_j}(B_i^j w-\bar{B}w)^T(B_i^j w-\bar{B}w)$$

$$= w^T\Big(\sum_{j=1}^{h}\sum_{i=1}^{n_j}(B_i^j-\bar{B})^T(B_i^j-\bar{B})\Big)w.$$

By defining $S_T \equiv \sum_{j=1}^{h}\sum_{i=1}^{n_j}(B_i^j-\bar{B})^T(B_i^j-\bar{B})$ as the image covariance matrix, 2D-PCA finds w^* such as

$$w^* = \arg max_w\, w^T S_T w \tag{3}$$

and the solution is obtained by solving the eigenvalue problem $S_T w=\lambda w$. The eigenvectors $w_1, w_2, \cdots, w_l$ corresponding to the largest eigenvalues of $S_T w=\lambda w$ are used to produce a reduced dimensional representation of A_i^j such as

$$(A_i^j)^* \equiv [B_i^j w_1, \cdots, B_i^j w_l].$$

The distance between the transformed representations of two images $A_{i_1}^{j_1}$ and $A_{i_2}^{j_2}$ can be computed by

$$distance((A_{i_1}^{j_1})^*,(A_{i_2}^{j_2})^*) = \sum_{k=1}^{l}\| B_{i_1}^{j_1}w_k - B_{i_2}^{j_2}w_k\|_2. \tag{4}$$

For other possible distance measures, one can refer to [6].

In [5] where 2D-PCA was proposed, partitioning of a face image was limited to a column-type, i.e., a window size was fixed as $1\times c$. It was generalized to any size of rectangles in [6] where 2D-LDA was introduced. 2D-LDA finds projective directions which maximize between-class distances and minimize within-class scatters. It is formulated by using the between-class scatter matrix S_B and the within-class scatter matrix S_W such as

$$\arg max_w \frac{w^T S_B w}{w^T S_W w}, where \tag{5}$$

$$S_B = \sum_{j=1}^{h} n_j (B^{(j)}-\bar{B})^T(B^{(j)}-\bar{B}), \tag{6}$$

$$S_W = \sum_{j=1}^{h}\sum_{i=1}^{n_j}(B_i^j - B^{(j)})^T(B_i^j - B^{(j)}), \tag{7}$$

and $B^{(j)}=\frac{1}{n_j}\sum_{i=1}^{n_j}B_i^j$ is the mean of class j with the elements $B_1^j, \cdots, B_{n_j}^j$. Eq. (5) is computed by solving the generalized eigenvalue problem $S_B w=\lambda S_W w$ to obtain the eigenvectors corresponding to the largest eigenvalues.

Since a height of the scatter matrices S_B and S_W is $\frac{r}{s}\times\frac{c}{t}$, S_B and S_W are usually nonsingular under reasonable window sizes and an eigenvalue problem in 2D-dimension reduction is formed by much smaller scatter matrices than in traditional PCA or LDA. The performance of 2D-PCA and 2D-LDA has been shown to be superior to 1D-PCA or 1D-LDA in face recognition. However, unlike LDA where the reduced dimension is usually set as the number of classes minus one, in 2D-LDA there is no specific rule to determine the optimal number of projective vectors. In most of literatures, the ratio of the sum of the largest eigenvalues to the total sum of eigenvalues has been used to determine the number of projective vectors, which is usually taken in the range of 95%~98%. However, as will be shown in the next section, it is not stable to use the ratio of eigenvalues in various data sets. In the next section, we present a practical method to determine the number of projective vectors and an ensemble approach in 2D-LDA.

3. AN ENSEMBLE CLASSIFIER IN 2D-LDA

3.1 On determining the number of projective vectors

The eigenvectors $w_1, w_2, \cdots, w_l$ corresponding to the largest eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_l$ in the generalized eigenvalue problem $S_B w=\lambda S_W w$ give a transformed representation

$$[B_i^j w_1, \cdots, B_i^j w_l]$$

of a face image A_i^j. The number l of projective vectors makes great effects on the classification performance. Usually, the number l is determined

by computing the ratio of eigenvalues such as

$$\tau = \frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{\frac{r}{s}\times\frac{c}{t}} \lambda_i} \quad (8)$$

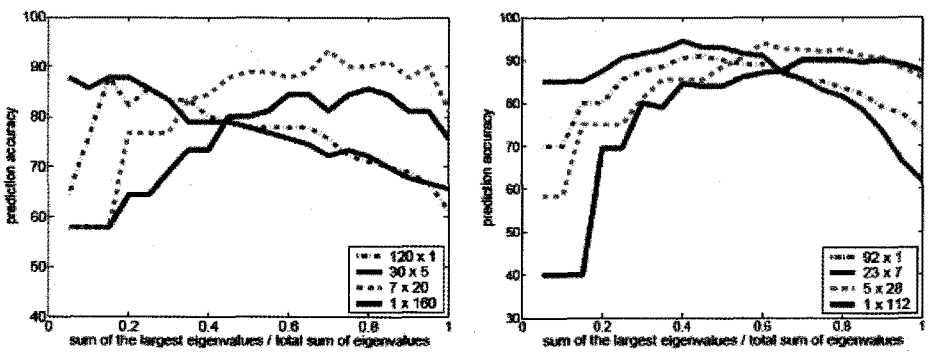
where λ_i's are in decreasing order and $\frac{r}{s}\times\frac{c}{t}$ is the total number of eigenvalues. However, it is not easy to determine l optimally by using the ratio τ in real data sets.

We performed preliminary experiments using Yale and AT&T face databases, whose goal is to measure prediction accuracies by changing the ratio τ in (8). Yale face database contains 165 images, 11 images of 15 subjects [7]. In our experiment, a face portion in each face image was manually cropped to a size of 120×160. The AT&T database has 400 images, which consists of 10 images of 40 subjects [8]. The size of each image is 92×112. In each database, the first 5 samples per subject were used as training data and the remaining as test data. Several window sizes were tested in 2D-LDA. Two figures in Figure 2 show prediction accuracies with respect to various τ values and window sizes in Yale and AT&T face database respectively. The range of τ which produced high prediction accuracies was varied depending on data and window sizes. The use of the eigenvectors corresponding to about 90% eigenvalues was not optimal contrary to the most of results in the literature [6,9,10].

In order to determine the number of projective vectors, k-cross validation(CV) can be used. CV is performed by dividing the training data as k parts equally. One of the parts is held out as a validation set and training is performed by using all the remaining elements, and it is repeated until all the parts are used once as a validation set. When the number of data samples is small as in face recognition problems, CV is usually unstable since a face image is high dimensional and the number of face images is much smaller than the data dimension. In that case, the leave-one-out method (LOO) can be used which is a special case of cross-validation. In LOO, each data sample is left out for validation and the remaining data is used for training. It is repeated by turns until every data sample is used as a validation sample. It means that eigenvalue problems should be solved for each hold-out case repeatedly and it can make the process expensive.

We propose a new approach for determining the number of projective vectors in 2D-LDA. The new method simplifies the leave-one-out procedure, resulting in saving computational complexities, but produces comparable or superior performances to the leave-one-out method as will be shown in the experiments conducted in Section 4.

We define the discriminative power of each pro-



(a) Yale (b) AT&T

Fig. 2. Comparison of prediction accuracies with respect to various τ values.

jective vector w_i as classification performance in the transformed feature space by w_i. The larger the eigenvalue is, the greater the discriminative power of the corresponding eigenvector is generally. However, it is not reliable to use eigenvalues directly as a measure of the discriminative power. Instead, we adopt a variation of leave-one-out procedure. The discriminative power of the projective vector w_i is measured by prediction accuracies in the transformed space by w_i. In the transformed space, a class label of each data sample is predicted by a 1-nearest neighbor classifier using the remaining training samples and it is repeated until all data samples are used as a leaved-out sample. The number of data samples whose class labels are predicted correctly is used for the discriminative power of the projective vector w_i, denoted as p_i. When $l+1$ is the smallest number such that $p_{l+1} < C \times p_1$ for some constant C, l eigenvectors are chosen for projection. This approach is summarized in Algorithm 1.

The difference between Algorithm 1 and classical leave-one-out method is that in classical leave-one-out method the eigenvalue problem $S_B w = \lambda S_W w$ should be solved in each case where one data sample is leaved out. On the other hand, in Algorithm 1, the eigenvalue problem is solved out only once using all the training samples, therefore saving computational complexities greatly.

3.2 An ensemble classifier

While the number of projective vectors is an important factor in dimension reduction, the size of a window for image partitioning also makes great impacts on 2D-LDA. In Section 3.1, we discussed on determining the number of projective vectors. In this section, as a tool for determining the number of projective vectors in any fixed window size, we consider Algorithm 1 and leave-one-out method (LOO).

One general way to choose a window size for 2D-LDA is to test several window sizes while determining the optimal number of projective vectors for each window size by LOO and select the one which gives the best average prediction accuracy on the training data. Let us call this approach as LOO-Selection.

On the other hand, an ensemble method can be used to combine several sizes of windows instead of selecting one model. Ensemble approaches have been used in data mining and machine learning by aggregating the predictions of multiple classifiers [11,12,13]. The family for ensemble can be composed in various ways. In our experiments, an ensemble family is composed of various types of windows from a row-type window of a size $r \times 1$ to a column-type window of a size $1 \times c$. Starting from a size $r \times 1$, a width is halved and a height is doubled each time until reaching a window size

Algorithm 1. the procedure to determine the number of projective vectors in 2D-LDA

Given training image samples $B_1, B_2, \cdots, B_n$ and a threshold constant C such that $0 < C < 1$,

1. Solve the eigenvalue problem $S_B w = \lambda S_W w$ where S_B and S_W are constructed by Eq. (6).
2. Let w_i be an eigenvector corresponding to the eigenvalue λ_i where λ_i's are sorted in decreasing order.
3. For each w_i,
4. Construct the feature space transformed by w_i, $\{b_1 = B_1 w_i, b_2 = B_2 w_i, \cdots, b_n = B_n w_i\}$.
5. Predict a class label for each b_j, $j = 1, \cdots, n$, by 1-NN classifier based on the other data samples $\{b_1, \cdots, b_{j-1}, b_{j+1}, \cdots, b_n\}$.
6. Compute the discriminative power p_i of the projective vector w_i by the number of data samples whose class labels are predicted correctly.
7. If $p_i < C \times p_1 (i \neq 1)$, then set $l = i - 1$ and go out from the for-loop.
8. end for
9. Obtain the projective vectors $\{w_1, w_2, \cdots, w_l\}$.

$1 \times c$. For each window size, projective vectors are computed by 2D-LDA, the number of which is determined by Algorithm 1 or LOO. Original data space is transformed by projective vectors and a classifier is modeled in the transformed data space. The ensemble family is constructed by the classifiers and voting scheme is used for the final decision. We call this approach as Algo1-Ensemble or LOO-Ensemble. Figure 3 summarizes the proposed method.

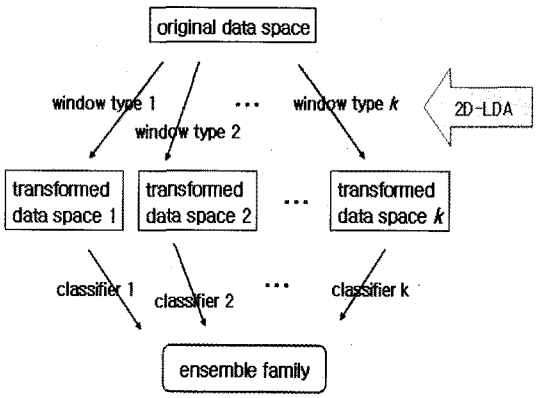


Fig. 3. Description of the proposed method.

4. EXPERIMENTAL RESULT

Using Yale and AT&T face databases, we compare performance of several methods, LOO-Selection, Bayes-Selection, LOO-Ensemble and Algo1-Ensemble including traditional 1D-LDA. For 1D-LDA, since the scatter matrices in LDA become singular due to high data dimension and a small number of samples, preprocessing by PCA was used for dimension reduction before applying LDA. The reduced dimension by PCA was set to the number of training samples minus one. Bayes-Selection uses Bayes errors presented in [6]. For Algorithm 1, two values 0.5 and 0.7 were used for the threshold C.

For an ensemble family, the window sizes $\{120\times1, 60\times2, 30\times5, 15\times10, 7\times20, 3\times40, 1\times80, 1\times160\}$ for Yale database and $\{92\times1, 46\times3, 23\times7, 11\times14, 5\times28, 2\times56, 1\times112\}$ for AT&T database were used. The 1-NN classifier was employed for classification in the dimension reduced space.

Table 1 compares prediction accuracies. The

Table 1. Comparison of prediction accuracies (%)

Yale		k=2	k=3	k=4	k=5
1D-LDA		65.2	81.7	82.9	83.3
2D-LDA	Bayes-Selection	79.3	75.8	81.9	80.0
	LOO-Selection	79.3	76.7	82.9	86.7
	LOO-Ensemble	79.3	81.7	84.8	93.3
	Algo1(0.5)-Ensemble	74.8	82.5	84.8	87.8
	Algo1(0.7)-Ensemble	78.5	83.3	86.7	91.1
AT&T		k=2	k=3	k=4	k=5
1D-LDA		77.8	84.3	86.7	81.5
2D-LDA	Bayes-Selection	78.1	82.1	85.4	84.0
	LOO-Selection	83.1	91.1	91.3	86.0
	LOO-Ensemble	87.8	90.4	93.3	91.5
	Algo1(0.5)-Ensemble	87.2	90.4	92.9	93.5
	Algo1(0.7)-Ensemble	88.4	90.7	91.7	92.0

number k in each column means that the first k face images of each subject were used for training data. The highest prediction accuracies in each column are denoted as a bold face. As shown in Table 1, prediction accuracies by Algo1-Ensemble and LOO-Ensemble are much higher compared with other methods in most of cases. Also note that Algorithm 1 needs to solve an eigenvalue problem only once, while in a classical leave-one-out method an eigenvalue problem should be solved as many times as the number of training data samples. Hence the combination of Algo1 and ensemble approaches can give good performance in both prediction accuracies and computational savings.

5. DISCUSSION

We proposed an ensemble approach for 2D-LDA which can overcome parameter selection problems and give competent classification performance in face recognition. In order to determine the number of projective vectors for feature extraction, a variant of a leave-one-out method is proposed reducing computational complexity and finding near optimal solution. An ensemble family utilizing various window sizes in 2D-LDA is constructed. The experimental results using face databases demonstrate the superiority of the proposed methods. Except one case among eight cases, ensemble approaches gave better prediction accuracies than the approach of selecting one window size. While in all the cases the proposed method *Algo1-Ensemble* showed competent performance to *LOO-Ensemble*, the computational complexity of *Algo1-Ensemble* can be much lower than that of *LOO-Ensemble*.

REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces v.s. Fisherfaces: Recognition using class specific linear projection," *IEEE transactions on pattern analysis and machine learning*, Vol.19, No.7, pp. 711-720, 1997.

[2] P. Howland and H. Park, *A comprehensive survey of text mining*, chapter: Cluster-preserving dimension reduction methods for efficient classification of text data, pp. 3-23, Springer-Verlag, 2003.

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

[4] J. Yang and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?," *Pattern recognition*, Vol.36, pp. 563-566, 2003.

[5] J. Yang, D. Zhang, F. Grangi, and J. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, Vol.26, No.1, pp. 131-137, 2004.

[6] C. Kim and C. Choi, "Image covariance-based subspace method for face recognition," *Pattern recognition*, Vol.40, pp.1592-1604, 2007.

[7] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[8] http://www.uk.research.att.com/facedatabase.html.

[9] J. Yang, J. Y. D. Zhang, and B. Niu, "Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics," *IEEE transactions on pattern analysis and machine intelligence*, Vol.29, No.4, pp. 650-664, 2007.

[10] D. Hu, G. Feng, and Z. Zho, "Two-dimensional locality preserving projections (2dLPP) with its application to palmprint recognition," *Pattern recognition*, Vol.40, pp. 339-342, 2007.

[11] P. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, Addison Wesley, Boston, MA, 2006.

[12] T. Dietterich, "Ensemble methods in machine

learning," In the proceedings of the first international workshop on multiple classifier systems, 2000.

[13] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, Vol.55, No.1, pp. 119-139, 1997.

Cheong Hee Park

1991. 2. Yonsei University, Dept. of Mathematics, B.S.
1998. 2. Yonsei University, Dept. of Mathematics, Ph.D.
2004. 8. University of Minnesota, Dept. of computer science and engineering, Ph.D.
2005. 4~Chungnam national university, Dept. of computer science and engineering, Assistant professor.
Research interests: Pattern recognition, Data mining