FACE VERIFICATION USING LOCALLY LINEAR DISCRIMINANT MODELS

Marios Kyperountas, Anastasios Tefas [1], Ioannis Pitas

Aristotle University of Thessaloniki, Dept. of Informatics, GR-54124, Thessaloniki, Greece
[1] Dept. of Information Management, Technological Institution of Kavala, GR-65404, Kavala, Greece

ABSTRACT

When linear discriminant analysis (LDA) is employed, the correct classification of a sample heavily depends on having an adequately large training set. This is often not possible in practical applications, such as person verification, where the lack of sufficient training samples causes improper estimation of a linear separation hyper-plane between the two classes. To overcome this shortcoming a novel algorithm that can handle the verification problem more efficiently than traditional LDA is presented. The dimensionality of the samples is reduced by breaking them down, thus creating subsets of smaller dimensionality feature vectors, and applying discriminant analysis on each subset. The resulting discriminant weight sets are themselves weighted under a normalization criterion, making the discriminant functions continuous in this sense. A series of simulations that formulate the face verification problem illustrate the cases for which our method outperforms traditional LDA and various statistical observations are made about the discriminant coefficients that are generated.

Index Terms— discriminant analysis, face verification, small sample size problem

1. INTRODUCTION

Linear discriminant analysis is an important statistical tool for recognition, verification, and in general classification applications. In many cases, however, and in particular when face data is used, there is insufficient data available so as to carry out the LDA process in a statistically proper manner. In face verification systems a test face is compared against a reference face and a decision is made whether the test face is identical to the reference face (meaning the test face is a client) or not (meaning the test face is an impostor). In this type of problems, Fisher's linear discriminant is not expected to be able to discriminate well between face pattern distributions that are in many cases highly nonlinear (i.e. they cannot be separated linearly), unless a sufficiently large training set is available. More specifically, in face recognition or verification systems

LDA-based approaches often suffer from the "*small sample size*" (SSS) problem where the dimensionality of the samples is larger than the number of training samples [1]. In fact, when this problem becomes severe, traditional LDA actually degrades the classification performance and shows poor generalization ability.

In recent years, an increasing interest has developed in the research community in order to improve LDA-based methods and provide solutions for the SSS problem. The traditional solution to this problem is to apply LDA in a lower-dimensional PCA subspace, so as to discard the null space (i.e., the subspace defined by the eigenvectors that correspond to zero eigenvalues) of the within-class scatter matrix of the training data set [2]. However, it has been shown [3] that significant discriminant information is contained in the discarded space and alternative solutions have been sought. Specifically, in [4] a direct-LDA algorithm is presented that discards the null space of the between-class scatter matrix, which is claimed to contain no useful information, rather than discard the null space of the within-class scatter matrix. In [5], a linear feature extraction method which is capable of deriving discriminatory information of the LDA criterion in singular cases is used. This is a two-stage method, where PCA is first used to reduce the dimensionality of the original space and then a Fisher-based linear algorithm, called Optimal Fisher Linear Discriminant, finds the best linear discriminant features on the PCA subspace. The authors in [1] form a mixture of LDA models that can be used to address the high nonlinearity in face pattern distributions, a problem that is commonly encountered in complex face recognition tasks. They present a machine-learning technique that is able to boost an ensemble of weak learners slightly better than random guessing to a more accurate learner. One of the major disadvantages of using the Fisher criterion is that the number of its discriminating vectors capable to be found is equal to the number of classes minus one. Recently, it was shown [6] that alternative LDA schemes that give more than one discriminative dimensions, in a two class problem, have better classification performance than those that give one projection.

This paper presents and evaluates an algorithm that aims to improve the performance of LDA-based approaches under the verification, or two-class, problem. The dimensionality of the samples is reduced by breaking them down and creating subsets of feature vectors with small dimensionality, and applying discriminant analysis on each subset. The resulting discriminant weights are normalized to provide the overall solution. This process gives direct improvements to the two aforementioned

problems as the non-linearity between the data pattern distributions is now restricted while the reduced dimensionality also helps mend the SSS problem. The performance of this method is studied using randomly generated data for various two-class problems that fall under various cases of the SSS problem.

2. LOCALLY LINEAR DISCRIMINANT MODELS

Let $\mathbf{m}_{r,C}$ and $\mathbf{m}_{r,I}$ denote the sample mean of the class of similarity vectors $\mathbf{c}_t$ that corresponds to client claims relating to the reference person r (intra-class mean) and those corresponding to impostor claims relating to person r (inter-class mean), respectively. In addition, let N_C and N_I be the corresponding numbers of similarity vectors that belong to these two classes and N be their sum, i.e., the total number of similarity vectors. Let $\mathbf{S}_W$ and $\mathbf{S}_B$ be the within-class and between-class scatter matrices, respectively [7]. Suppose that we would like to transform linearly the similarity vectors:

$$D'(t,r) = \mathbf{w}_r^{\mathrm{T}} \mathbf{c}_t . \qquad (1)$$

The most known and plausible criterion is to find a projection, i.e. choose $\mathbf{w}_r$, that maximizes the ratio of the between-class scatter against the within-class scatter (Fisher's criterion):

$$J(\mathbf{w}_r) = \frac{\mathbf{w}_r^{\mathrm{T}} \mathbf{S}_B \mathbf{w}_r}{\mathbf{w}_r^{\mathrm{T}} \mathbf{S}_W \mathbf{w}_r} . \qquad (2)$$

For the two-class problem, as is the case of face verification, Fisher's linear discriminant provides the vector that maximizes (2) and is given by:

$$\mathbf{w}_{r,0} = \mathbf{S}_W^{-1}(\mathbf{m}_{r,I} - \mathbf{m}_{r,C}) . \qquad (3)$$

When the number of client similarity vectors N_C is smaller than the dimensionality L of each vector $\mathbf{c}_t$ then traditional LDA shows poor generalization ability and degrades the classification performance since the SSS problem appears.

The novelty of our approach is that in order to give remedy to the SSS problem, each similarity vector $\mathbf{c}_t'$ with dimensionality L' is broken down to P smaller dimensionality vectors, $\mathbf{c}_{t,i}'$, $i = 1, \ldots P$, each one of length M, where $M \le (N_C - 1)$, thus uniformly forming P subsets. As a result, P separate Fisher linear discriminant processes are carried out and each of the weight vectors produced is normalized, so that the within group variance equals to one, by applying:

$$\mathbf{w}_{r,0,i}' = \mathbf{w}_{r,0,i} (\mathbf{w}_{r,0,i}^{\mathrm{T}} \mathbf{S}_{W,i} \mathbf{w}_{r,0,i})^{-\frac{1}{2}} , \qquad (4)$$

where $i = 1, \ldots, P$ is the index of M – dimensionality vector $\mathbf{c}_{t,i}'$ corresponding to a subset of similarity vector coordinates. This normalization step enables the proper merging of all weight vectors to a single column weight vector, $\mathbf{w}_{r,0}'$, as such:

$$\mathbf{w}_{r,0}' = \left[\mathbf{w}_{r,0,1}'^{\mathrm{T}}, \ldots, \mathbf{w}_{r,0,P}'^{\mathrm{T}} \right]^{\mathrm{T}} . \qquad (5)$$

3. EVALUATION STUDY OF ALGORITHM

The efficiency of the proposed discriminant solution is evaluated using simulated data in order to deduce experimental evidence on the performance of WPLDA. In order to provide relevant background on the expected performance of the proposed WPLDA algorithm in face verification, simulations that tackle the 2-class problem are carried out. We intent to investigate the cases where one can expect the WPLDA algorithm to outperform the traditional LDA algorithm, with respect to the size of the impostor and client classes. For each verification experiment, two classes of matching vectors, one that corresponds to the clients and the other to the impostors, are created. Each class contains N sample vectors of dimensionality L. Each of these sample vectors contains entries drawn from a normal (Gaussian) distribution. The L random entries to each sample vector of class X_j, which is the $j-$th client or impostor class, are generated by

$$N_{X_j}\left(x_i^j : \mu_i^j, \sigma_i^j\right) = \frac{1}{\sigma_i^j \sqrt{2\pi}} e^{-\frac{\left(x_i^j - \mu_i^j\right)^2}{2\left(\sigma_i^j\right)^2}} , i = 1 \ldots L , \quad (6)$$

where $\mu_i^j = G_j + \alpha r_i$ and $\sigma_i^j = K_j + \beta r_i$. G_j is the expected mean value and K_j the expected standard deviation for the $i-$th random entry of the $j-$th class and r_i is a random number, chosen from a normal distribution with zero mean and unit variance. The scalars α and β affect the uniformity among the vectors of each class.

The dimensionality of the sample vectors is set to $L = 64$ and each class contains $N = 2000$ sample vectors. Let I be the impostor class and C_1 and C_2 be two client classes. Let the random entries to each sample vector of the impostor class I and the client classes C_1 and C_2 be generated based on the following normal distributions, respectively:

$$N_I\left(x_i : \mu_i = 100 + 5r_i, \sigma_i = 25 + 5r_i\right), i = 1 \ldots 64. \qquad (7)$$

$$N_{C_1}\left(x_i : \mu_i = 87 + 5r_i, \sigma_i = 35 + 5r_i\right), i = 1 \ldots 64. \qquad (8)$$

$$N_{C_2}\left(x_i : \mu_i = 85 + 5r_i, \sigma_i = 35 + 5r_i\right), i = 1 \ldots 64. \qquad (9)$$

It is clear that the mean of the random entries of C_2 is expected to deviate more, w.r.t. the mean of the entries of C_1, from the mean of the entries of I.

For most feature-based verification methods it is expected that certain matching features should provide more discriminant information than others. For example, for the case of face verification, a feature related to the eye is expected to be more useful than a feature related to a part of the forehead. In order to simulate a similar situation, we create a subset of L_B characteristics (out of the total L), that is expected to be more discriminant than the remaining characteristics. We name this set of L_B characteristics as 'most discriminant coefficients'. Let a client class C_3 be created, such that the entries at the L_B characteristics are taken from the C_2 client class (since the entries from C_2 are more separated from the entries in I than
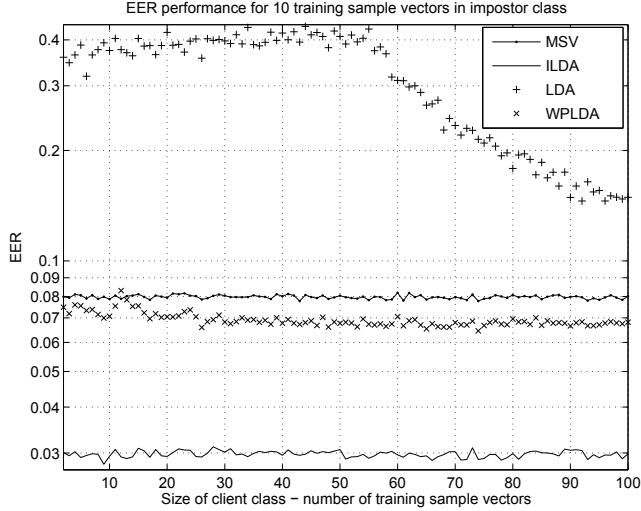
Fig. 1. *EER* when varying the number of client sample vectors selected from class C_3 and for 10 impostor training vectors selected from class I.
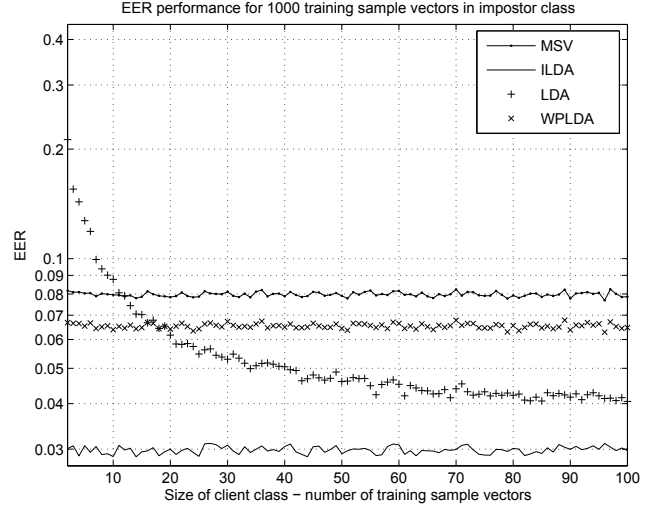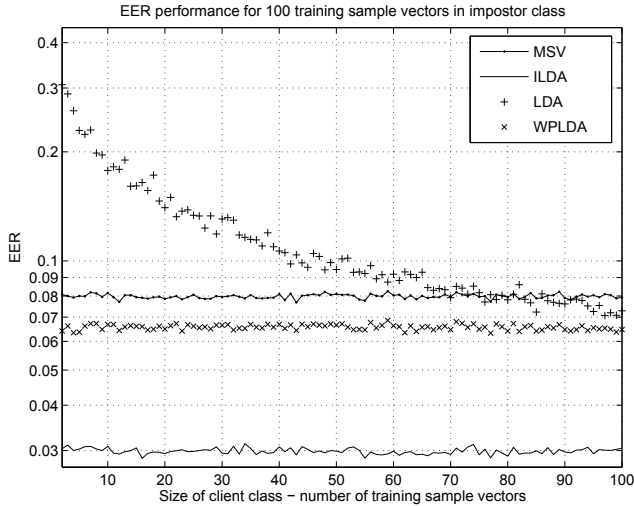

Fig. 2. *EER* when varying the number of client sample vectors selected from class C_3 and for 100 impostor training vectors selected from class I.

the entries of C_1 are) and the rest of the entries from the C_1 class. For this first set of experiments we let $L_B = 5$.

The data that were created are used to compare the discrimination ability of traditional LDA and the proposed WPLDA for various numbers of training sample vectors for the impostor and client class. For each 2-class problem that is formulated, one training and one test set are created. The training set of LDA and WPLDA is formed based on the random selection out of the complete set of N sample vectors of each class. The remaining sample vectors of each class, obtained by excluding the training set of LDA and WPLDA, form the test set that is used to evaluate the classification performance.

In order to approximate the ideal linear discriminant solution, a third method that will be referred to as Ideal LDA (ILDA) will

Fig. 3. *EER* when varying the number of client sample vectors selected from class C_3 and for 1000 impostor training vectors selected from class I.

always apply the traditional LDA algorithm making use of the complete sets of N client sample vectors and N impostor sample vectors, during the training phase. The test set, where the performance of ILDA will be evaluated on, is identical to the test set of LDA and WPLDA. Thus, the test set is always included in the training set of ILDA, so as to best approximate the ideal linear discriminant solution and provide ground-truth results. In addition, and again for comparison purposes, the classification performance of a fourth method will be considered, where this method simply computes the mean of the sample vectors (MSV) and produces a non-weighted result which can be used to indicate how difficult the 2-class classification problem is.

In order to evaluate the performance of the four aforementioned methods the equal error rate (*EER*) is found over 20 independent runs, for more accurate results, and the average value is recorded. The simulation data are used in various discriminant processes that aim to separate out the client and impostor classes. The 2-class problem that is studied next uses data from I and C_3. Fig. 1-3 show the *EER* when the number of client sample vectors varies from 2 to 100.

Fig. 1 shows the EER results when the number of impostor sample vectors is 10. For the LDA algorithm, the SSS is expected to have the most severe effects on the EER when the client class has less than $(L+1) = 65$ samples. In theory, in this case neither the client class nor the impostor class can be properly modelled by traditional LDA and, as a result, an appropriate separation between the two classes cannot be found. On the other hand, WPLDA is not affected by the SSS problem as can be seen in Fig. 1. The small variations in the *EER* of ILDA indicate the amount of randomness in our results since only the *y*-axis showing *EER* is significant for the ILDA results. Fig. 2 and 3 show the *EER* rates for 100 and 1000 impostor sample vectors respectively. It is clearly seen in these figures that, unless a relatively large number of client and impostor sample vectors are available, WPLDA outperforms LDA.

Fig. 2 shows that, when 100 impostor and 83 client sample vectors are available, the performance of LDA becomes better than that of MSV. Fig. 3 shows that when the number of impostor sample vectors becomes 1000, 20 client sample vectors are required for LDA to outperform WPLDA. For most current biometric databases, having 20, or more, client samples per person is quite uncommon.

The second set of experiments using simulated data involves investigating the statistical behaviour of the discriminant coefficients of the LDA and WPLDA processes with reference to ILDA. Moreover, *EER* rates are reported for different numbers of 'most discriminant coefficients' contained in each class, that is, for various values of L_B. In order to determine how efficient each discriminant method is in recognizing the importance of the most discriminant coefficients, a separation criterion between the most discriminant and the remaining coefficients is defined as:

$$H = \frac{|m_B - m_R|}{s_B + s_R}, \qquad (10)$$

where m_B and s_B are scalars representing the average mean and the average standard deviation of the set of most discriminant coefficients and m_R and s_R those of the remaining coefficients. If $H \geq 1$, the separation criterion is satisfied since the values of the most discriminant coefficients will vary significantly from those of the remaining coefficients.

This set of simulations is modelled under the SSS problem, where the client class has less sample vectors than the dimensionality of the similarity vectors. Both the *Brussels* protocol, described in [7], and the *Lausanne* [8] protocol, which are widely used for verification purposes, specify for the client class to avail 6 training samples during the training stage. Therefore, we randomly select 6 sample vectors from the C_3 client class and 1000 sample vectors from the I impostor class to train LDA and WPLDA. The coefficients of ILDA are once again generated by a training set of 2000 client and 2000 impostor sample vectors. To observe the statistical behaviour of the discriminant coefficients, 1000 independent runs are carried out. The entries at the position of the L_B elements are expected to have a larger distance from the corresponding element entries of class I, than the rest. Thus, the discriminant process should give larger weights for the element entries at these L_B specific positions, since they are expected to be the most useful in producing a meaningful separation between the impostor and the client class.

Table 1 provides statistical information about the calculation of the 64 discriminant coefficients, $\mathbf{w}_i^{'}$, $i = 1\ldots 64$, throughout the 1000 independent runs, by ILDA, LDA and WPLDA respectively. The three discriminant methods processed the C_3, for various values of L_B, and I training data. The H values in Table 1 show that WPLDA provides a better separation for the 'most discriminant coefficients' from the remaining coefficients, in terms of assigning largest weights, than LDA does. In addition, WPLDA provides the *EER* rate that is closest to the

TABLE I
MEAN *EER* AND MEAN *H*

L_B	*EER* ILDA	*EER* LDA	*EER* WPLDA	*H* ILDA	*H* LDA	*H* WPLDA
0	0.0961	0.2578	**0.0989**	-	-	-
5	0.0300	0.1142	**0.0652**	3.7908	0.4580	**4.5505**
15	0.0255	0.1032	**0.0318**	1.0739	0.1321	**4.7119**
25	0.0247	0.1008	**0.0272**	0.3986	0.0489	**1.6225**
30	0.0246	0.1014	**0.0267**	0.1532	0.0252	**0.1703**

Mean *EER* and *H* of 1000 runs for 6 clients and 1000 impostors.

corresponding ILDA rate.

4. CONCLUSION

A novel methodology is proposed in this paper that provides general solutions for LDA-based algorithms that encounter problems relating to high nonlinearity between the data pattern distributions, small training sets and to the SSS problem in particular. This methodology was evaluated using a set of simulations that gave indications on when the proposed weighted piecewise linear discriminant analysis algorithm outperforms traditional LDA. It is anticipated that the performance of other LDA variants may be enhanced by utilizing processes that stem from this framework.

5. REFERENCES

[1] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Boosting linear discriminant analysis for face recognition", in *Proc. IEEE Int. Conf. on Image Processing (ICIP'03),* vol. 1, pp. I - 657-60, Barcelona, Spain, 14-17 Sep., 2003.

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

[3] L-F Chen, *et al.*, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, vol. 33, pp. 1713–1726, 2000.

[4] 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.

[5] J. Yang and J. Yang, "Why can LDA be performed in PCA transformed space?", *Pattern Recognition*, vol. 36, pp. 563-566, 2003.

[6] C. Songcan and Y. Xubing, "Alternative linear discriminant classifier", *Pattern Recognition*, vol. 37, no. 7, pp. 1545-47, 2004.

[7] C. Kotropoulos, A. Tefas and I. Pitas, "Frontal face authentication using morphological elastic graph matching", *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp. 555-560, April 2000.

[8] J. Luettin, and G. Maitre, "Evaluation protocol for the extended M2VTS database (XM2VTSDB)," in *IDIAP Communication 98-05*, IDIAP, Martigny, Switzerland, 1998.