MOTIVATING CLASS-SPECIFIC NONLINEAR PROJECTIONS FOR FACE VERIFICATION

Georgios Goudelis, Stefanos Zafeiriou, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle university of Thessaloniki, Greece

ABSTRACT

In this paper we motivate the use of class-specific nonlinear subspace methods for face verification. The problem of face verification is considered as a two-class problem (genuine versus impostor class). The typical *Fisher's Linear Discriminant Analysis* (FLDA) gives only one or two projections in a two-class problem. This is a very strict limitation to the search of discriminant dimensions. As for the FLDA for N class problems ($N > 2$) the transformation is not person specific. In order to remedy these limitations of FLDA, exploit the individuality of human faces and take into consideration the fact that the distribution of facial images, under different viewpoints, illumination variations and facial expression is highly complex and non-linear, novel kernel discriminant algorithms are used. The new method is tested in the face verification problem using various datasets where it is verified that it outperforms other commonly used kernel approaches

Index Terms— Face verification, two-class problems, kernel techniques, Fisher's linear discriminant analysis.

1. INTRODUCTION

It is widely accepted that the distribution of facial images, under different viewpoints, illumination variations and facial expression is highly complex and non-linear [1, 2, 3]. Thus, a variety of nonlinear techniques has been developed in order to successfully capture the underlying nonlinearity of data and the most popular have been the so-called kernel techniques [1, 2, 3]. The two problems of face verification and recognition are conceptual different and should be treated differently when extracting discriminant features for treating them. On one hand face recognition is treated as a multiclass problem, where the space is separated to various face classes. On the other hand the strategy for face verification is to find class-specific projections that separate the genuine (client) class from the impostor class.

Moreover, there differences in the measures that are used for evaluating the performance of both face verification and recognition. In many cases for evaluating the performance of a face recognition system, only the percentage of correctly identified faces within a number of matches is adequate (recognition rate) [1, 2]. On the other hand the performance of face verification systems is measured in terms of the *False Rejection Rate* (FRR) achieved at a fixed *False Acceptance Rate* (FAR) [4]. There the trade-off between FAR and FRR creates a curve where FRR is plotted as a function of FAR. The performance of a verification system is often quoted by a particular operating point of this curve where FAR=FRR [4]. This operating point is called *Equal Error Rate* (EER). The EER will be used to quantify the performance of the tested methods, in the experimental results section.

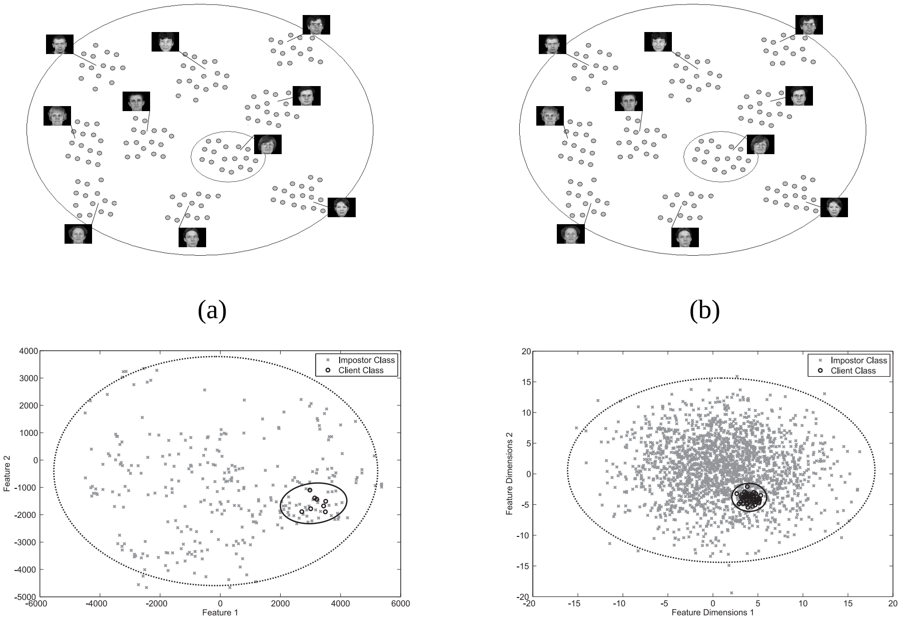
2. NONLINEAR CLASS-SPECIFIC DISCRIMINANT FEATURE EXTRACTION

In order to make use of kernel techniques the original input space is projected to an arbitrary-dimensional space $\mathcal{F}$ (the space $\mathcal{F}$ usually has the structure of a Hilbert space [2]). To do so, let $\phi : \mathbf{z} \in \Re^M \longrightarrow \phi(\mathbf{z}) \in \mathcal{F}$ be a nonlinear mapping from the input space $\Re^M$ to the Hilbert space $\mathcal{F}$.

Let r be the reference person that will be used for defining the person specific algorithms. The genuine vectors $\mathbf{z}_i \in \Re^M$ of the person r will be denoted as $\boldsymbol{\rho}_i = \mathbf{z}_i$ ($\mathbf{z}_i \in \mathcal{U}_r$), while the impostor images $\mathbf{z}_i$ of the person r will be denoted as $\boldsymbol{\kappa}_i = \mathbf{z}_i$ ($\mathbf{z}_i \in \mathcal{I}_r$). Let also $\bar{\boldsymbol{\rho}}$, $\bar{\boldsymbol{\kappa}}$ and $\bar{\mathbf{m}}$ be the mean vectors of the genuine class, the impostor class and total mean of the facial vectors in the Hilbert space $\mathcal{F}$. Any function k satisfying the Mercer's condition can be used as a kernel. The dot product of $\phi(\mathbf{z}_i)$ and $\phi(\mathbf{z}_j)$ in the Hilbert space can be calculated without having to evaluate explicitly the mapping $\phi(\cdot)$ as $k(\mathbf{z}_i, \mathbf{z}_j) = \phi(\mathbf{z}_i)^T \phi(\mathbf{z}_j)$ (this is also known as the kernel trick). The typical kernels that have been used have been polynomial and Radial Basis Functions (RBF). Kernels that do not satisfy the Mercer's condition have been successfully applied for face recognition [3] [i.e., Fractional Polynomial Models (FPM)] and have been considered in the experiments.

The criterion that is used in this paper, will be formed using a simple similarity measure in the Hilbert space $\mathcal{F}$. This measure quantifies the similarity of a given feature vector $\mathbf{z}$ to the reference facial class r in the subspace spanned by the columns of the matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \dots \boldsymbol{\psi}_K]$, with $\boldsymbol{\psi}_i \in \mathcal{F}$. The L_2 norm in the reduced space spanned by the columns of $\boldsymbol{\Psi}$, is used as similarity measure:

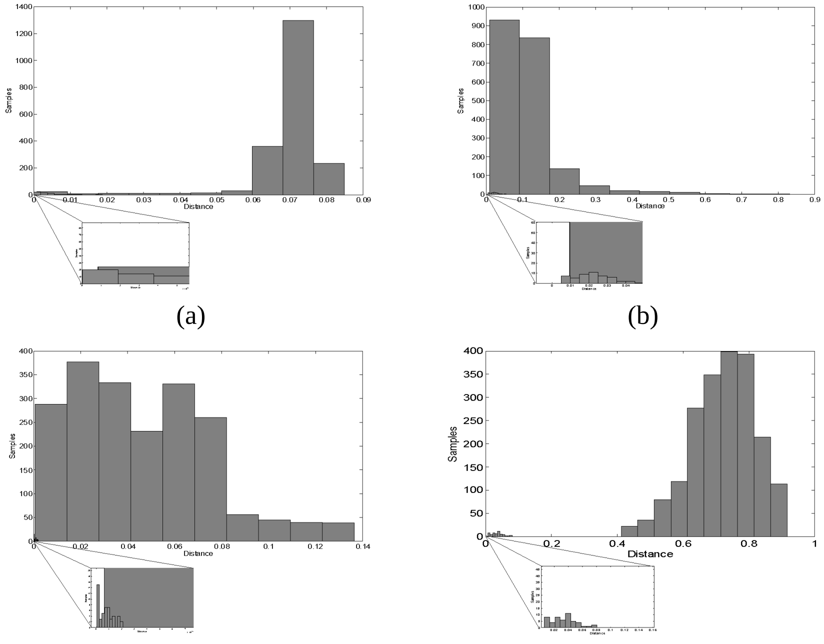
$$\begin{aligned} d_r(\mathbf{z}) &= ||\boldsymbol{\Psi}^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})||^2 \\ &= \mathrm{tr}[\boldsymbol{\Psi}^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T \boldsymbol{\Psi}] \\ &= \textstyle\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T \boldsymbol{\psi}_i] \end{aligned} \quad (1)$$



(a) (b)

(c) (d)

Fig. 1. a) Multiclass face recognition modelling; b) two Class face verification modelling; c) the distribution of the first two features projected to the first two principal components; d) a simulation distribution derived from two bivariate normal distributions for impostors and clients



(a) (b)

(c) (d)

Fig. 2. Histograms of sample distances with: a)kernel Fisher’s discriminant analysis; b)kernel Fisher’s discriminant analysis with more than one dimensions by adding the a small noisy diagonal matrix to the between class scatter matrix; c) proposed kernel discriminant analysis with only the first dimension; d) proposed kernel discriminant analysis with 100 dimensions.

which is actually the Euclidean distance of a projected sample to the projected mean of the reference class and is one of most usually employed measures in pattern recognition applications (i.e, distance from the center of the class). This distance should be low for the samples of the genuine class and should be high for the samples of the impostor class.

Now, in order to find a discriminant linear transformation in $\mathcal{F}$ we demand that the sum of the similarity measures $d_r(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{I}_r$ (impostor similarity measures) to be maximized while minimizing the sum of the similarity measures $d_r(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{U}_r$ (client similarity measures). Thus, the discriminant projections $\boldsymbol{\psi}_i \in \mathcal{F}$ are found in the training set as the ones that maximize the ratio:

$$\begin{aligned} D^{\Phi}(\mathbf{\Psi}) &= \frac{\sum_{\mathbf{z}\in\mathcal{I}_r} d_r(\mathbf{z})}{\sum_{\mathbf{z}\in\mathcal{U}_r} d_r(\mathbf{z})} \\ &= \frac{\sum_{\mathbf{z}\in\mathcal{I}_r}\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T\boldsymbol{\psi}_i]}{\sum_{\mathbf{z}\in\mathcal{U}_r}\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T\boldsymbol{\psi}_i]} \\ &= \frac{\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T[\sum_{\mathbf{z}\in\mathcal{I}_r}(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T]\boldsymbol{\psi}_i]}{\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T[\sum_{\mathbf{z}\in\mathcal{U}_r}(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T]\boldsymbol{\psi}_i]} \\ &= \frac{\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T\mathbf{W}^{\Phi}\boldsymbol{\psi}_i]}{\sum_{i=1}^{K} \mathrm{tr}[\boldsymbol{\psi}_i^T\mathbf{B}^{\Phi}\boldsymbol{\psi}_i]} \\ &= \frac{\mathrm{tr}[\mathbf{\Psi}^T\mathbf{W}^{\Phi}\mathbf{\Psi}]}{\mathrm{tr}[\mathbf{\Psi}^T\mathbf{B}^{\Phi}\mathbf{\Psi}]}. \end{aligned} \tag{2}$$

where $\mathbf{W}^{\Phi} = \sum_{\mathbf{z}\in\mathcal{I}_r}(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T$, $\mathbf{B}^{\Phi} = \sum_{\mathbf{z}\in\mathcal{U}_r}(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T$ and $\mathrm{tr}[\mathbf{M}]$ is the trace of matrix $\mathbf{M}$. Direct optimization of $D^{\Phi}(\mathbf{\Psi})$ in $\mathcal{F}$ is an intractable problem due to the fact that both $\mathbf{W}^{\Phi}$ and $\mathbf{B}^{\Phi}$ are matrices with arbitrary dimensions. Thus, in order to extract features by the above criterion methods similar to [1, 2] have been used. The propose discriminant analysis will be called Class-Specific Kernel Discriminant Analysis (CSKDA) in the rest of the paper.

3. MULTICLASS VERSUS TWO-CLASS MODELLING

In Figures 1a and b the two different modelings (i.e, face recognition and face verification) can be seen. An example of two class face verification problem for 39 persons from the XM2VTS database, is illustrated in Figure 1c. For every person the first two features, derived from the projection to the two dominant eigenvectors of PCA (Principal Component Analysis), are depicted.

A simulation example can be found in Figure 1d where two classes have been created using bivariate normal distributions. The first class represents the client class, having 50 samples, while the second one models the impostor class, containing 2000 samples. It is obvious that non-linear methods should be applied in order to capture the distribution of the data. In order to provide some first insights of the benefits of CSKDA, we have applied non-linear modelling using RBF kernels in the artificial data of Figure 1d. The kernel Fisher discriminant alternatives give a very limited subspace of one dimension [1, 2, 5]. On the other hand Kernel PCA [1, 2, 5] provides a set of features, but has the disadvantage that does not consider class distribution characteristics. For the simulation example in Figure 1d the KPCA resulted in an 100 dimensional space. The proposed approach has resulted in an 100 dimensional space, as well.

Let that the similarity between a data sample, in the new space, and the genuine class, be measured using the Euclidean distance to center of the genuine class. The distribution of the client and impostor similarities, after applying KFDA and CSKDA, with the client class can be found in Figure 2. The zoomed area represents the distribution of the client distances. As can be seen, in Figure 2a, when using the one dimensional space of KFDA the data are somewhat well separable. When more dimensions are kept by adding an diagonal matrix with small noisy elements to the between-class scatter matrix, the two classes are heavily confused (see Figure 2b).

In many cases, in approximation and regularization theory [5, 3], a scaled version of the identity matrix is added to a matrix in order to become invertible [5, 3]. The scaled version of the identity matrix is a simplified version of the noisy diagonal matrix that we have used in the experiments. Using this fact, we provide a theoretical indication concerning why the use of additional dimensions of between class scatter matrix deteriorates the performance. As can be proven the matrix (in the two class case) [2] has only one eigenvector that corresponds to non null eigenvector. Let that we diminish the null eigenvalues of $\mathbf{S}_b^{\Phi}$ by adding the scaled version of the identity matrix as:

$$\mathbf{S}_b^{\Phi}\boldsymbol{\zeta} = 0 \Leftrightarrow \mathbf{S}_b^{\Phi}\boldsymbol{\zeta} + \sigma\boldsymbol{\zeta} = \boldsymbol{\zeta} \Leftrightarrow (\mathbf{S}_b^{\Phi} + \sigma\mathbf{I})\boldsymbol{\zeta} = \sigma\boldsymbol{\zeta}, \tag{3}$$

where $\sigma > 0$. Thus, the eigenvectors of $\mathbf{S}_b^{\Phi}$ that correspond to null eigenvalues are the same ones that correspond to eigenvalues equal to σ for the matrix $\mathbf{S}_b^{\Phi} + \sigma\mathbf{I}$. The property of the projection to the null eigenvectors of $\mathbf{S}_b^{\Phi}$, that may indicates poor classification performance is that if for some $\boldsymbol{\zeta} \in \mathcal{H}$, $\boldsymbol{\zeta}^T\mathbf{S}_b^{\Phi}\boldsymbol{\zeta} = 0$ then under the projection $\boldsymbol{\zeta}$, for the two training mean vectors (genuine and impostor) in feature space $\mathcal{H}$, it is valid that, $\boldsymbol{\zeta}^T\bar{\boldsymbol{\rho}} = \boldsymbol{\zeta}^T\bar{\boldsymbol{\kappa}}$. In other words under the projection $\boldsymbol{\zeta}$ the two centers, $\bar{\boldsymbol{\rho}}, \bar{\boldsymbol{\kappa}}$ fall in the same point, which means that this projection does not help in separating the two classes (is not optimal in sense of FLDA, where this projection make the criterion equals to zero).

On the other hand the samples of the two classes are not well separated using only the first dimension of the proposed method, but they become fully separated when using 100 dimensions. Let that the maximum distance of the client samples be considered as a threshold for accepting or rejecting a claim (this means that false rejection equals to zero). Using this threshold, in Figure 3, a comparison of false acceptances introduced from KFDA, KPCA and the proposed techniques for various kept dimensions, is shown. As can be seen when more than one dimensions are kept for KFDA, by adding the identity matrix to the between class scatter matrix, the performance deteriorates and more false acceptances are introduced. On the other hand the performance of KPCA and the

Table 1. A Comparison of the best EERs measured using Gabor feature vectors and fractional polynomial models at various feature extraction methods

Algorithm	EER%-XM2VTS	EER%-ORL	EER%-Yale
Gabor + KPCA with Fractional Polynomial Models	10.2	5.3	7
Gabor + Best of Multiclass KFDA [1, 2, 3] with Fractional Polynomial Models	6.8	4.2	3.4
Gabor + CSKDA with Fractional Polynomial Models	3.3	3.2	1.6

proposed kernel technique increases with the number of kept dimensions.

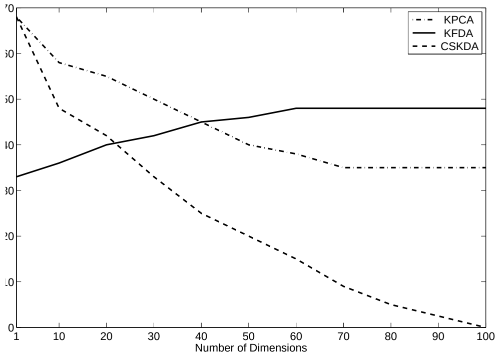


Fig. 3. Number of false acceptances versus the dimensionality.

4. EXPERIMENTS WITH REAL DATA

The databases that have been used in our experiments have been ORL, Yale and the XM2VTS databases. For these databases in order to make maximal use of the data we have considering a circular protocol. In order to implement this protocol, we have combined principles of the leave one out strategy and the rotation estimates, i.e., a variant of the jack-knife method [4]. In each circle of the protocol one of the persons become the impostor and its images are used for impostor claims (not seen in the training phase). Then, the 80% of the data of the remaining persons are used for training and the remaining 20% serve for client claims.

Gabor-based facial features combined with kernel methods (e.g., KPCA and variants of multiclass kernel Fisher's discriminant analysis [3]) and with fractional polynomial models are among the state-of-the-art face verification and recognition systems in the literature. We have conducted experiments using the augmented Gabor features proposed in [3]. Moreover, we have applied the proposed method using these Gabor features and we have verified that it has superior performance and outperforms Gabor-KPCA and multiclass Gabor-KFDA with FPM models. The best EERs in the various tested databases are summarized in Table 1.

5. CONCLUSION

Face verification has been modelled as a nonlinear two class problem (clients vs impostors). The majority of discriminant feature extraction methods that are used for face recognition, are based on Fisher's discriminant analysis. The analysis in this paper indicates that: 1)the one dimensional space of two class KFDA may be insufficient for correctly representing data in two class cases, 2) simple tricks, like adding noisy diagonal matrices to the between class scatter matrix, in order to have larger KFDA spaces, deteriorate the performance and 3) the proposed criterion provides a multidimensional space where the data can be well represented. Moreover our method has been tested in face verification using various face databases, where they show to outperform many other popular kernel methods.

6. REFERENCES

[1] L. Juwei et. al., "Face recognition using kernel direct discriminant analysis algorithms," *IEEE T-NN*, vol. 14, no. 1, pp. 117–126, 2003.

[2] J. Yang et. al., "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE T-PAMI*, vol. 27, no. 2, pp. 230–244, 2005.

[3] L. Chengjun, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE T-PAMI*, vol. 28, no. 5, pp. 725–737, 2006.

[4] A. Tefas et. al., "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE T-PAMI*, vol. 23, no. 7, pp. 735–746, 2001.

[5] S. Mika et. al., "Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces," *IEEE T-PAMI*, vol. 25, no. 5, pp. 623 – 628, 2003.