Biometric Verification by Projections in Error Subspaces

Mariusz Leszczynski and Władysław Skarbek

Warsaw University of Technology, Faculty of Electronics and Information Technology,
00-665 Warszawa, Nowowiejska 15/19, Poland
M.Leszczynski@ire.pw.edu.pl

Abstract. A general methodology for design of biometric verification system is presented. It is based on linear feature discrimination using sequential compositions of several types of feature vector transformations: data centering , orthogonal projection onto linear subspace, vector component scaling, and orthogonal projection onto unit sphere. Projections refer to subspaces in global, within-class, and between-class error spaces. Twelve basic discrimination schemes are identified by compositions of subspace projections interleaved by scaling operations and single projection onto unit sphere. For the proposed discriminant features, the Euclidean norm of difference between query and average personal feature vectors is compared with the threshold corresponding to the required false acceptance rate. Moreover, the aggregation by geometric mean of distances in two schemes leads to better verification results. The methodology is tested and illustrated for the verification system based on facial 2D images.

Keywords: biometrics, face verification, discriminant analysis, singular subspace, within-class errors.

1 Introduction

Biometrics is a research field with a practical goal: create applications for uniquely recognizing humans based upon one or more intrinsic physical and/or behavioral traits including facial 2D/3D image, voice, fingerprints, eye retinas and irises, hand measurements, signature, gait and typing patterns. Biometric verification is one of three tasks which are usually attributed to pattern recognition: object identification, object verification, and similar object searching. However, biometric pattern verification is conceptually different from traditional class membership verification. To understand this point let us consider two pattern verification queries:

1. Given an image of a digit, verify whether the digit is *five*.
2. Given a facial image and a person identifier, verify whether the image matches to this id.

To solve the first problem a model for image class *five* is designed and used to verify the membership of the input image to the queried class. For instance symbol images x are mapped into a space of features $y = \mathcal{M}(x)$ in which memberships to symbol classes are represented by class c probability distributions

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 166–173, 2007.

$p_c(y)$ Then the predicate $\forall c \neq 5, p_c(y) < p_5(y)$ could be the basis of the verification. Moreover, such verification is optimal since it results in minimum of verification error = false acceptance rate + false rejection rate.

To solve the second problem we may follow the above approach. But then each new human being h in the system should have a new model p_h for his/her facial images in certain feature space. It means that models built for facial databases in training stage cannot be directly used in testing and exploiting stages of such verification system since in practice the sets of *training persons* and *exploiting persons* are different.

From the above examples we see that for the biometric verification we need such a model training procedure which builds a model with parameters to be used by testing and exploiting procedures.

Since natural human centered pattern classes cannot be used in person verification biometric systems, another categorization has to be sought. It appears that differences of human features for the biometric measurements of the same person (within-class differences) and for different persons (between-class features) create a consistent categorization including two specific classes. The specificity of this two classes follows from the fact that means of these two classes are both equal to zero. Moreover, for the within-class feature variation could be sometimes greater than between-class feature variation, i.e. usually the squared within-class errors are of the same magnitude as squared between-class errors.

Therefore, it is natural to look for such a linear transformation $W : \mathbb{R}^N \to \mathbb{R}^n$ of original biometric measurements $x \in \mathbb{R}^N$ (e.g. vectorized pixel matrix of face image or its 2D frequency representation) into a target feature vector $y = W^t x$ for which within-class differences are decreased while between-class differences are increased.

To this goal the class separation measure is defined as the ratio of between-class variation to within-class variation for vectorial data set $\{x_1, \ldots, x_L\}$ represented in columns of matrix $X \in \mathbb{R}^{N \times L}$:

$$v_X := \frac{\text{variation}_b(X)}{\text{variation}_w(X)} \tag{1}$$

where the within and between class variations are defined together with total variation via squared Euclidean distance:

$$\begin{aligned}
\text{variation}_w(X) &:= \frac{1}{J^2} \sum_{j=1}^{J} \frac{1}{L_j^2} \sum_{i_1, i_2 \in I_j} \|y_{i_1} - y_{i_2}\|^2 \\
\text{variation}_b(X) &:= \frac{1}{J^2} \sum_{j_1 \neq j_2} \frac{1}{L_{j_1} L_{j_2}} \sum_{i_1 \in I_{j_1}, i_2 \in I_{j_2}} \|y_{i_1} - y_{i_2}\|^2 \\
\text{variation}_t(X) &:= \text{variation}_w(X) + \text{variation}_b(X) = \\
&\frac{1}{J^2} \sum_{j_1=1}^{J} \sum_{j_2=1}^{J} \frac{1}{L_{j_1} L_{j_2}} \sum_{i_1 \in I_{j_1}, i_2 \in I_{j_2}} \|y_{i_1} - y_{i_2}\|^2
\end{aligned} \tag{2}$$

where J is the number of classes, L_j is the number of j-th class samples ($L = L_1 + \cdots + L_J$) whose index set is denoted by I_j, $j = 1, \ldots, J$.

It appears that the class variations are not new concepts as they are scaled forms of class variances which were introduced by Fisher already in thirties of twentieth century [1]:

$$\begin{aligned}
&\mathrm{var}_w(X) := \tfrac{1}{J}\sum_{j=1}^{J}\tfrac{1}{L_j}\sum_{i\in I_j}\|x_i-\overline{x}^j\|^2\\
&\mathrm{var}_b(X) := \tfrac{1}{J}\sum_{j=1}^{J}\|\overline{x}^j-\overline{x}\|^2\\
&\mathrm{var}_t(X) := \mathrm{var}_w(X)+\mathrm{var}_b(X) = \tfrac{1}{J}\sum_{j=1}^{J}\tfrac{1}{L_j}\sum_{i\in I_j}\|x_i-\overline{x}\|^2
\end{aligned} \tag{3}$$

where $\overline{x}^j$ is the class mean of all j-th class samples in X and $\overline{x}$ is the grand mean of all samples in X.

Namely, the following relations are true for class variations and class variances:

$$\begin{aligned}
&\mathrm{variation}_w(X) = \tfrac{2\mathrm{var}_t(X)}{J}\\
&\mathrm{variation}_b(X) = 2\left(\mathrm{var}_b(X)+\tfrac{J-1}{J}\mathrm{var}_w(X)\right)\\
&\mathrm{variation}_t(X) = 2\mathrm{var}_t(X)
\end{aligned} \tag{4}$$

Hence, the class separation measure v_X is the affine form of Fisher separation measure with coefficients solely dependent on the number of classes J :

$$v_X = Jf_X + J - 1 \tag{5}$$

2 Classical Optimization of Fisher Measure

The Fisher class separation measure becomes a goal function w.r.t. transformation matrix $W \in \mathbb{R}^{N\times n}$ when the source data matrix X is replaced by feature data matrix $Y := W^tX$.

The standard approach in optimizing (maximizing) $f(W) := f_{W^tX}$ is replacing the scalar product of two vectors by the trace of their outer product:

$$a^tb = \mathrm{tr}(ab^t),\ \|a\|^2 = \mathrm{tr}(aa^t)$$

Then we observe that the within and between-class variances are traces of within and between-class covariance matrices, respectively:

$$\begin{aligned}
&R_w(Y) := \tfrac{1}{L}\sum_{j=1}^{J}\tfrac{1}{L_j}\sum_{i\in I_j}(y_i-\overline{y}^j)(y_i-\overline{y}^j)^t = W^tR_w(X)W\\
&R_b(Y) = \tfrac{1}{J}\sum_{j=1}^{J}(\overline{y}^j-\overline{y})(\overline{y}^j-\overline{y})^t = W^tR_b(X)W\\
&f(W) = f_Y = \tfrac{\mathrm{tr}(R_b(Y))}{\mathrm{tr}(R_w(Y))} = \tfrac{\mathrm{tr}(W^tR_b(X)W)}{\mathrm{tr}(W^tR_w(X)W))}
\end{aligned} \tag{6}$$

The results of optimization for $f(W)$ are traditionally called Linear Discriminant Analysis (LDA). Fisher considered the scalar LDA features, i.e. the case of $n = 1$ in which $W = w \in \mathbb{R}^{n\times 1}$, $y = w^tx$ is the scalar and the within and between-class variances are quadratic forms of vectorial variable w. Then the Fisher measure transforms to Rayleigh quotient w.r.t. matrices R_b and R_w :

$$f(w) = f_y = f_{w^tX} = \frac{w^tR_b(X)w}{w^tR_w(X)w} \tag{7}$$

The standard analysis of stationary points for $f(W)$, $W = [w_1, \ldots, w_n]$, $w_i \in \mathbb{R}^N, i = 1, \ldots, n$, leads to conclusion that the maximum is achieved by eigenvectors w_i corresponding to the maximal eigenvalue λ_{max} of the following generalized eigenvalue problem:

$$R_b(X)W = \lambda R_w(X)W \tag{8}$$

Therefore the rank of matrix W cannot be higher than the rank of eigenvalue λ_{max}. In practice this rank equals to one and we get the result equivalent to scalar case with $n = 1$. Therefore, the additional requirement should be imposed onto $W : \mathrm{rank}(W) = n$.

If the matrix R_w is not singular then the optimal solution (Fukunaga [2]) at this requirement is achieved from Eigenvalue Decomposition (EVD) of symmetric, semi-definite matrix $R'_b := C_w^{-1} R_b C_w^{-t}$, where C_w is the Cholesky matrix ([3]) for R_w. Firstly, we look for W of rank N as follows:

$$\begin{array}{l} R_b W = \lambda R_w W,\ R_w = C_w C_w^t,\ W' = C_w^t W \\ R'_b = W' \Lambda (W')^t \\ W = C_w^{-t} W' \end{array} \tag{9}$$

If columns of W' are sorted by decreasing eigenvalues λ_i then we select from W the first n columns as the solution. This procedure works only if $\mathrm{rank}(R_w) = N$ and $\mathrm{rank}(R_b) \geq n$.

In Section 3 we discuss the important case of $\mathrm{rank}(R_w) < N$.

3 Optimization of Fisher Measures by Projections in Error Spaces

In case of singular matrix R_w a sort of regularization is necessary. There are known two general approaches to this problem:

1. Regularization of data by mapping to $Y = \mathcal{P}(X)$ in order to get nonsingular $R_w(Y)$.
2. Regularization of LDA model by imposing an additional constraint on full rank LDA matrix $W = [w_1, \ldots, w_n]$ – for instance orthogonality to kernel space of R_w :

$$w_i \perp \ker(R_w),\ i = 1, \ldots, n \tag{10}$$

In this section a novel point of view on LDA regularization is presented which uses the concept of projections in error subspaces. It unifies in one consistent scheme both approaches and integrates also with Dual Linear Discriminant Analysis (DLDA) [5].

In the presented discriminant analysis the source data matrix $Y_0 := X$ undergoes up to seven linear transformations before reaching the final matrix of features:

$$Y_{t-1} \longrightarrow Y_t,\ t = 1, \ldots, T \leq 7$$

The j-th class indexes I_j are identified by column indexes of data matrix Y_t and they are not changed at data matrix transformations.

There are three types of errors in our approach. They are defined w.r.t. any data matrix $Y = [y_1, \ldots, y_L]$ and with the fixed class assignments I_j, $j = 1, \ldots, J$:

1. *Grand error:* the difference of data vector y_k and the grand mean vector of Y. This can be modelled by the *global centering* operation $\mathcal{C}_g$:

$$\bar{y} = \frac{1}{L}\sum_{i=1}^{L} y_i,\ \mathcal{C}_g(y_k) := y_k - \bar{y},\ k = 1, \ldots, L \tag{11}$$

2. *Within-class error:* the difference of data vector $y_k, k \in I_j$ and its class mean $\bar{y}^{(j)}$:

$$\bar{y}^{(j)} = \frac{1}{L_j}\sum_{i \in I_j} y_i,\ \mathcal{C}_w(y_k) := y_k - \bar{y}^{(j)},\ k \in I_j,\ j = 1, \ldots, J \tag{12}$$

3. *Between-class error:* the difference of class mean $\bar{y}^{(j)}$ and grand mean $\bar{y}$:

$$\mathcal{C}_b(\bar{y}^{(j)}) := \bar{y}^{(j)} - \bar{y},\ j = 1, \ldots, J \tag{13}$$

The error vectors span error linear subspaces denoted as follows:

$$\mathcal{E}_g(Y) := \mathrm{span}(\mathcal{C}_g(Y)),\ \mathcal{E}_w(Y) := \mathrm{span}(\mathcal{C}_w(Y)),\ \mathcal{E}_b(\overline{Y}) := \mathrm{span}(\mathcal{C}_b(\overline{Y})) \tag{14}$$

Note that $\mathcal{E}_g(Y)$ is related to famous PCA approach recently linked to rough set [4] verification, too.

The singular bases $U^{(g)}$, $U^{(w)}$, $U^{(b)}$ of the error linear subspaces are obtained from Singular Value Decomposition (SVD [3]) for matrices $\mathcal{C}_g(Y), \mathcal{C}_w(Y), \mathcal{C}_b(\overline{Y})$, respectively:

$$\mathcal{C}_g(Y) = U^{(g)}\Sigma^{(g)}(V^{(g)})^t,\ \mathcal{C}_w(Y) = U^{(w)}\Sigma^{(w)}(V^{(w)})^t,\ \mathcal{C}_b(\overline{Y}) = U^{(b)}\Sigma^{(b)}(V^{(b)})^t$$

where diagonal squared matrices $\Sigma^{(\cdot)}$ are of size equal to the rank of centered data matrix $\mathcal{C}_\cdot()$. In case of grand and within-class centering singular values are ordered from maximal to minimal value while in case of between-class centering the standard SVD order is inverse – the first element on the diagonal is minimal.

Let $Y \in \mathbb{R}^{a \times L}$. Then we identify all singular subspaces of dimension $a' \leq \dim(\mathcal{E}(Y))$ of error spaces by the projection operators which map the space $\mathbb{R}^a$ onto $\mathbb{R}^{a'}$ – the space of projection coefficients w.r.t. to the singular base $U_{a'}$ restricted to the first a' vectors:

1. $\mathcal{P}^{(g)}_{a,a'}$: projection onto grand error singular subspace of dimension a';
2. $\mathcal{P}^{(w)}_{a,a'}$: projection onto within-class error singular subspace of dimension a';
3. $\mathcal{P}^{(b)}_{a,a'}$: projection onto between-class error singular subspace of dimension a'.

Additional operation required after projection is component-wise scaling by the first inverse singular values which create the diagonal matrix $\Sigma_{a'}^{-1}$:

1. $\mathcal{S}_{a'}^{(g)}$: scaling of projected vector in grand error singular subspace of dimension a';
2. $\mathcal{S}_{a'}^{(w)}$: scaling of projected vector in within-class error singular subspace of dimension a';
3. $\mathcal{S}_{a'}^{(b)}$: scaling of projected vector in between-class error singular subspace of dimension a'.

In matrix composition terms the projection and scaling operations have the form:

$$\mathcal{P}_{a,a'}(x) = U_{a'}^t x,\ \mathcal{S}_{a'}(y) = \Sigma_{a'}^{-1} y \tag{15}$$

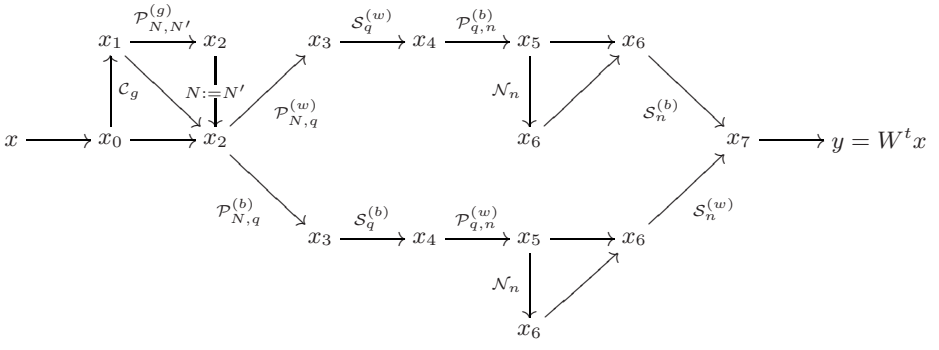
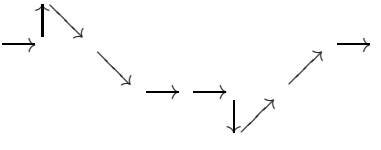


Fig. 1. Diagram of LDA type transformations based on projections onto error singular subspaces

The last operation we use in definition of LDA transformation is the vector length normalization $\mathcal{N}_a$ which can be geometrically interpreted as the projection on the unit sphere in $\mathbb{R}^a$.

$$\mathcal{N}_a(x) := \frac{x}{\|x\|} \tag{16}$$

Using the above notation the all known to authors LDA transformations can be defined via the diagram in Fig.1. It defines altogether 12 LDA type transformations. For instance in face verification system the following transformation path on the diagram gives best results:



In terms of operation compositions we get the following sequence (denoted here by DLDA) which includes also optimal weighting for matching:

$$\mathcal{W}_{DLDA} := \mathcal{S}_n^{(w)} \mathcal{N}_n \mathcal{P}_{q,n}^{(w)} \mathcal{S}_q^{(b)} \mathcal{P}_{N,q}^{(b)} \mathcal{C}_{(g)} \tag{17}$$

It is better than more popular LDA scheme improved by centering and normalization operations:

$$\mathcal{W}_{LDA} := \mathcal{S}_n^{(b)} \mathcal{N}_n \mathcal{P}_{q,n}^{(b)} \mathcal{S}_q^{(w)} \mathcal{P}_{N,q}^{(w)} \mathcal{C}_{(g)} \tag{18}$$

As a matter of fact the discriminant operation maximizing DLDA class separation measure is restricted to the composition $\mathcal{P}_{q,n}^{(w)} \mathcal{S}_q^{(b)} \mathcal{P}_{N,q}^{(b)} \mathcal{C}_{(g)}$ while the final two operations $\mathcal{S}_n^{(w)} \mathcal{N}_n$ are responsible for the optimal thresholding of within-class error which is selected as the distance function for the person id verification.

4 Experiments for Face Verification

From the previous works described in [5] it is already known that in case of face verification the optimization of inverse Fisher ratio (DLDA) leads to better results than the optimization of Fisher ratio (LDA). The final weighting of LDA or DLDA vector components had been also applied since they follow from Gaussian model of class errors.

Moreover, it was also observed that the normalization operation $\mathcal{N}_n$ improves significantly the equal error rate and ROC function face verification based on LDA or DLDA. The reason is explained by weak correlation between within-class error and between-class error. Therefore despite comparable norm magnitude of

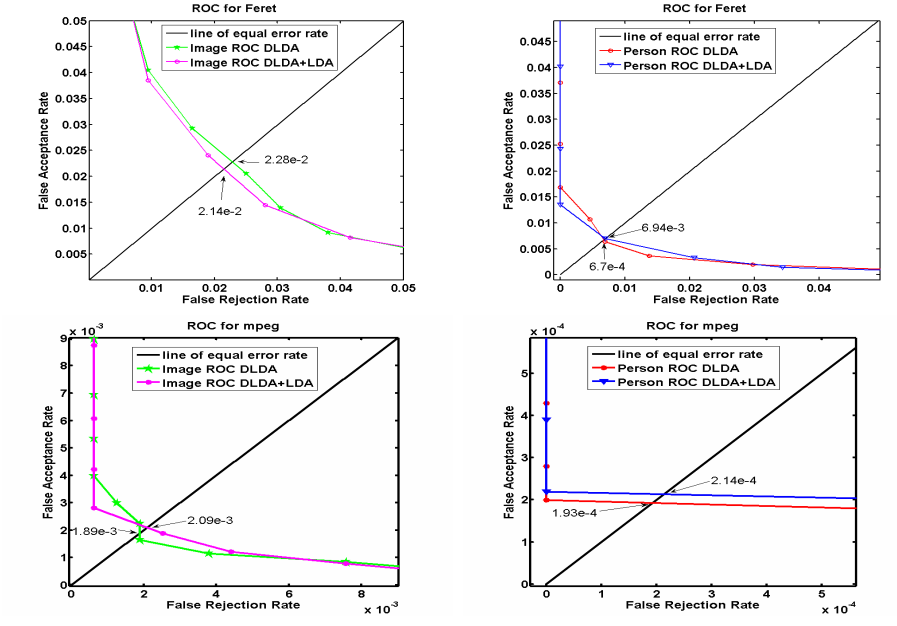


Fig. 2. Receiver operating characteristics and equal error rate for two facial databases *Feret* and *Mpeg* for DLDA and combined LDA+DLDA in single and multi-image scenarios

those errors the projection onto the unit sphere separates them while the final scaling respects the probability of all errors which are projected onto the same point of the unit sphere.

In the experiments described here we analyze two problems for face verification:

1. Is there any combination of LDA and DLDA class errors which improves DLDA?
2. What is the degree of improvement if the verification is based on the several facial images of the same person instead of the single one?

For the first problem we have found (cf. Fig. 2):

- LDA and DLDA class errors are of comparable magnitude.
- Geometric mean of both errors leads to slight improvements of EER and ROC w.r.t. DLDA error alone.
- The maximum, the arithmetic mean, and the harmonic mean of LDA and DLDA class errors give intermediate results between the best DLDA results and significantly worse LDA results.

For the second problem, experiments prove the significant advantage of the multi-image approach. Confront Fig. 2 where by *Person ROC* we mean the id acceptance if at least half of the query images are accepted. The acceptance by single image is described by *Image ROC*.

Conclusions. The general methodology presented for design of biometric verification system based on linear feature discrimination using sequential compositions of several types of feature vector transformations identifies the 12 basic discrimination schemes. The methodology has been tested for the verification system based on facial 2D images allowing for the choice of two best schemes which aggregated by geometric mean of distances leads to the best face verification results.

Acknowledgment. The work presented was developed within VISNET 2, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 Programme.

References

1. Fisher, R. A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7** (1936) 179–188
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1992)
3. Golub, G., Loan, C.: Matrix Computations. The Johns Hopkins University Press (1989)
4. Swiniarski, R.W., Skowron A.: Independent Component Analysis, Principal Component Analysis and Rough Sets in Face Recognition. In Transactions on Rough Sets I. Springer, LNCS **3100** (2004) 392-404
5. Skarbek, W., Kucharski, K., Bober, M.: Dual LDA for Face Recognition. Fundamenta Informaticae **61** (2004) 303–334