A New Covariance Estimate for Bayesian Classifiers in Biometric Recognition

Carlos E. Thomaz, Duncan F. Gillies, and Raul Q. Feitosa

***Abstract*—In many biometric pattern-recognition problems, the number of training examples per class is limited, and consequently the sample group covariance matrices often used in parametric and nonparametric Bayesian classifiers are poorly estimated or singular. Thus, a considerable amount of effort has been devoted to the design of other covariance estimators, for use in limited-sample and high-dimensional classification problems. In this paper, a new covariance estimate, called the maximum entropy covariance selection (MECS) method, is proposed. It is based on combining covariance matrices under the principle of maximum uncertainty. In order to evaluate the MECS effectiveness in biometric problems, experiments on face, facial expression, and fingerprint classification were carried out and compared with popular covariance estimates, including the reguralized discriminant analysis and leave-one-out covariance for the parametric classifier, and the Van Ness and Toeplitz covariance estimates for the nonparametric classifier. The results show that, in image recognition applications whenever the sample group covariance matrices are poorly estimated or ill posed, the MECS method is faster and usually more accurate than the aforementioned approaches in both parametric and nonparametric Bayesian classifiers.**

***Index Terms*—Bayesian classifiers, biometric recognition, covariance estimate, limited sample sizes, maximum entropy.**

I. Introduction

STATISTICAL pattern-recognition techniques have been used successfully to design several recognition systems [11]. In the statistical approach, a pattern is represented by a set of p features or parameters and the region of the feature space occupied by each class is determined by the probability distribution of its corresponding patterns, which must be either specified (parametric approach) or learned (nonparametric approach).

There are a number of classification rules available to define appropriate statistical decision-making boundaries [11]. The well-known Bayes' decision rule that assigns a pattern to the class with the highest posterior probability is the one that achieves minimal misclassification risk among all possible rules (see, e.g., [1]).

The idea behind the Bayes' rule is that all of the information available about class membership is contained in the set of conditional probability densities. In practice, most of these probability densities are based on Gaussian kernel functions that involve the inverse of the true covariance matrix of each class [3], [7], [8], [12], [13]. Since in real-world problems the true covariance matrices are seldom known, estimates must be computed based on the patterns available in a training set.

The usual choice for estimating the true covariance matrices is the maximum-likelihood estimate defined by the corresponding sample group covariance matrices. However, it is well known that in limited-sample-size applications the inverse of sample group covariance matrices is either poorly estimated or cannot be calculated when the number of training patterns per class is smaller than the number of features. This problem is indeed quite common nowadays, especially in image-recognition applications where patterns are frequently composed of thousands of pixels from which hundreds of preprocessed image features are obtained.

Thus, a considerable amount of effort has been devoted to the design of other covariance estimators, for targeting limited-sample and high-dimensional problems in parametric and nonparametric Bayesian classifiers [3]–[8], [17], [20], [23]. Most of these covariance estimators either rely on optimization techniques that are time consuming or have restrictive forms and do not necessarily lead to the highest classification accuracy in all circumstances.

In this paper, a new covariance estimate called maximum entropy covariance selection (MECS) is presented. This estimate is based on combining covariance matrices to take into account the maximum uncertainty. It assumes that there are some sources of variation that are the same from class to class and, consequently, similarities in covariance shape may be expected for all the classes. This has often been the case for biometric applications such as face recognition.

In order to evaluate the MECS effectiveness, experiments on face, facial expression, and fingerprint recognition were carried out, using publicly released databases with different ratios between the training sample sizes and the dimensionality of the feature space. In each application, the performance of the MECS approach was compared with other covariance estimators, including the reguralized discriminant analysis (RDA) and leave-one-out covariance (LOOC) for the parametric Bayesian classifier, and the Van Ness and Toeplitz covariance estimates for the nonparametric Bayesian classifier.

The results show that, in such recognition applications whenever the sample group covariance matrices were poorly estimated or ill posed, the MECS covariance method is preferable

Manuscript received October 18, 2002; revised May 7, 2003. The work of C. E. Thomaz was supported in part by the Brazilian Government Agency CAPES under Grant 1168/99-1.

C. E. Thomaz and D. F. Gillies are with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: cet@doc.ic.ac.uk; dfg@doc.ic.ac.uk).

R. Q. Feitosa is with the Department of Electrical Engineering, Catholic University of Rio de Janeiro, Rio de Janeiro 22453-900, Brazil, and also with the Department of Computer Engineering, State University of Rio de Janeiro, Rio de Janeiro 205590-900, Brazil (e-mail: raul@ele.puc-rio.br).

Digital Object Identifier 10.1109/TCSVT.2003.821984

to the other approaches in both parametric and nonparametric Bayesian classifiers. Moreover, the MECS approach offers considerable savings in computation time.

II. Limited-Sample-Size Problem

The similarity measures used for Bayesian classifiers based on Gaussian kernels involve the inverse of the true covariance matrices of each class.

Conventionally, those matrices are estimated by the sample group covariance matrices S_i, which are the unbiased maximum-likelihood estimators of the true covariance matrices [1], given by

$$S_i = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \tag{1}$$

where $x_{i,j}$ is the pattern j from class i and n_i is the number of training patterns from class i.

However, the inverse of S_i is especially problematic if for p-dimensional patterns less than $p+1$ training patterns from each class are available. Since the sample group covariance matrix is a function of $(n_i - 1)$ or less linearly independent vectors, its rank is $(n_i - 1)$ or less. Therefore, S_i is a singular matrix if n_i is less than the dimension of the feature space.

Another well-known problem related to the sample group covariance matrix is its instability due to limited samples. This effect can be explicitly seen by writing the S_i matrices in their spectral decomposition forms [4], that is,

$$S_i = \Phi_i \Lambda_i \Phi_i^T = \sum_{k=1}^{p} \lambda_{ik} \phi_{ik} \phi_{ik}^T \tag{2}$$

where λ_{ik} is the kth eigenvalue of S_i and ϕ_{ik} is the corresponding eigenvector. According to this representation, the inverse of the sample group covariance matrix can be calculated as

$$S_i^{-1} = \sum_{k=1}^{p} \frac{\phi_{ik}\phi_{ik}^T}{\lambda_{ik}}. \tag{3}$$

From (3), it can be observed that the inverse of S_i is heavily weighted by the smallest eigenvalues and the directions associated with their respective eigenvectors [3]. Hence, a poor or unreliable estimation of S_i tends to exaggerate the importance associated with the low-variance information and consequently distorts discriminant similarity measures based on these estimates.

As a general guideline, Jain and Chandrasekaran [10] have suggested that the class sample sizes should be at least five to ten times the dimension of the feature space. However, these estimation settings have been quite difficult to achieve in practice, particularly in biometric recognition problems where patterns are frequently composed of thousands of pixels or even hundreds of preprocessed image features.

In Section III, two of the most popular parametric and nonparametric Bayesian classifiers are briefly described along with their commonly used nonconventional covariance estimates for targeting limited sample and high dimensional problems. By "nonconventional covariance estimate," we mean any covariance estimate that is not based solely on the data of the sample group. A new covariance estimate that can be used in both types of Bayesian classifiers is detailed in Section V.

III. Quadratic Discriminant Classifier

The quadratic discriminant classifier is one of the most popular parametric Bayesian classifiers. It is based on the p-multivariate Gaussian class-conditional probability densities.

Assuming the symmetrical or zero-one loss function, the Bayes quadratic discriminant (QD) rule stipulates that an unknown pattern x should be assigned to the class i that *minimizes*

$$d_i(x) = \ln|S_i| + (x - \bar{x}_i)^T S_i^{-1} (x - \bar{x}_i) - 2\ln \pi_i \tag{4}$$

where π_i is a prior probability associated with the ith class, S_i is the maximum-likelihood estimate of the respective true covariance matrix (defined in (1)), and $\bar{x}_i$ is the maximum-likelihood estimate of the corresponding true mean vector, given by

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}. \tag{5}$$

The discriminant rule described in (4) defines the standard or conventional quadratic discriminant function (QDF) classifier.

A. Linear Discriminant Classifier

One straightforward method routinely applied to overcome the limited-sample-size problem on the QDF classifier, and consequently deal with the singularity and instability of the sample group covariance matrices S_i, is to employ the Fisher's linear discriminant function (LDF) classifier.

The LDF classifier is obtained by replacing the S_i in (4) with the pooled sample covariance matrix defined as

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{N - g} \tag{6}$$

where g is the number of classes and $N = n_1 + n_2 + \cdots + n_g$. Since more patterns are taken to calculate the pooled covariance matrix, S_p is indeed a weighting average of S_i, S_p will potentially have a higher rank than S_i and would be a full rank matrix.

Theoretically, however, S_p is a consistent estimator of the true covariance matrices Σ_i only when $\Sigma_1 = \Sigma_2 \cdots = \Sigma_g$.

B. RDA Estimate

In order to reduce the singularity and instability effects of the QDF classifier due to limited samples and the limitation of the LDF classifier, Friedman proposed an approach called the RDA [3].

RDA is a two-dimensional (2-D) optimization method that shrinks both the S_i toward S_p and also the eigenvalues of the S_i toward equality by blending the first shrinkage with multiples of the identity matrix.

In this context, the sample covariance matrices S_i of the QD rule defined in (4) are replaced by the following $S_i^{rda}(\lambda, \gamma)$:

$$\begin{aligned} S_i^{rda}(\lambda,\gamma) &= (1-\gamma)S_i^{rda}(\lambda) + \gamma\left(\frac{tr(S_i^{rda}(\lambda))}{p}\right) I \\ S_i^{rda}(\lambda) &= \frac{(1-\lambda)(n_i - 1)S_i + \lambda(N-g)S_p}{(1-\lambda)n_i + \lambda N} \end{aligned} \tag{7}$$

where the notation "tr" denotes the trace of a matrix. The mixing parameters λ and γ are restricted to the range 0–1 (optimization grid) and are selected to maximize the leave-one-out classification accuracy regarding all classes [3].

Although RDA has the benefit of being directly related to the classification accuracy, it is a computationally intensive method particularly when a large number of classes is considered.

C. LOOC Estimate

Hoffbeck and Landgrebe [8] proposed a covariance estimator for QDF classifiers that depends only on covariance optimization of single classes.

The idea is to examine pairwise mixtures of the sample group covariance estimates S_i and the unweighted common covariance estimate S, together with their diagonal forms [8]. The unweighted common covariance estimate S is given by

$$S = \frac{1}{g}\sum_{i=1}^{g} S_i \tag{8}$$

and can be viewed as the pooled covariance matrix when the number of training patterns is equal in each class.

The LOOC estimator has the following form:

$$S_i^{\mathrm{looc}}(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(S_i) + \alpha_i S_i, & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S, & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\mathrm{diag}(S), & 2 < \alpha_i \le 3. \end{cases} \tag{9}$$

Its optimization strategy consists of evaluating several values of α_i over the grid $0 \le \alpha_i \le 3$ and then choosing α_i that maximizes the average log likelihood of the corresponding p-dimensional Gaussian density function [8].

As the LOOC estimate requires that only one density function be evaluated for each point on the α_i optimization grid, it generally requires less computation than the RDA estimator.

IV. Parzen Window Classifier

The Parzen window classifier is a popular nonparametric Bayesian classifier. In this classifier, the class-conditional probability densities are estimated locally by using kernel functions and a number of group neighboring patterns [4].

Assuming the symmetrical or zero-one loss function, the Bayes classification rule stipulates that an unknown pattern x should be assigned to the class i corresponding to the highest or maximum posterior probability.

In the standard Parzen classifiers with Gaussian kernels (PZW), the posterior probability of each class is calculated by multiplying the prior probability of the ith class with the corresponding Parzen likelihood density estimate $q_i(x)$, given by

$$\begin{aligned} q_i(x) &= \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{(2\pi)^{p/2}} v_i(x - x_{i,j}) \\ v_i(x - x_{i,j}) &= \frac{1}{h_i^p\sqrt{|S_i|}} \\ &\quad \times \exp\left[-\frac{1}{2h_i^2}(x - x_{i,j})^T S_i^{-1}(x - x_{i,j})\right] \end{aligned} \tag{10}$$

where, as a reminder, p is the dimension of the feature space, $x_{i,j}$ is the pattern j from class i, n_i is the number of training patterns from class i, and S_i is the sample group covariance matrix. The parameter h_i is the window width of class i and controls the kernel function "spread" or size.

Due to the limited-sample-size problem, several researchers have imposed, analogously to the QDF classifiers, some structures on the sample group covariance matrices for use in Gaussian Parzen classifiers [4], [7], [12], [23]. Two approaches commonly employed for overcoming these estimation singularities and instabilities are briefly described in Sections IV-A and IV-B.

A. Van Ness Covariance Estimate

Van Ness proposed a flexible diagonal form for the true covariance matrices of Gaussian Parzen classifiers based solely on the estimation of the variances of each variable [23].

In this approach, the sample group covariance matrices of the Parzen density estimates defined in (10) are replaced with the following matrices:

$$S_i^{\mathrm{ness}}(\beta) = (\beta)\mathrm{diag}(S_i) \tag{11}$$

where β is a smooth or scale parameter selected to maximize the leave-one-out classification accuracy regarding all classes.

Since only the sample variance of each variable has to be calculated from the training patterns of each class, S_i^{ness} would be nonsingular as long as there are at least two linearly independent patterns available per class.

B. Toeplitz Covariance Estimate

Another possible structure for the covariance matrices is the Toeplitz approximation, based on the stationary assumption [4]. The basic idea is to allow each individual variable to have its own variance, whereas all covariance elements along any diagonal are multiplied by the same correlation factor.

The Toeplitz approximation of each group covariance matrix can be calculated as follows:

$$S_i^{\mathrm{toep}} = \Gamma_i P_i \Gamma_i \tag{12}$$

where

$$\Gamma_i = \begin{bmatrix} \sigma_{i1} & & & 0 \\ & \sigma_{i2} & & \\ & & \ddots & \\ 0 & & & \sigma_{ip} \end{bmatrix}$$

$$P_i = \begin{bmatrix} 1 & \rho_i & \cdots & \rho_i^{p-1} \\ \rho_i & 1 & & \vdots \\ \vdots & & \ddots & \rho_i \\ \rho_i^{p-1} & \cdots & \rho_i & 1 \end{bmatrix} \tag{13}$$

and σ_{ik} is the sample standard deviation of variable k calculated from the training patterns of class i. The correlation factor ρ_i is given by the average of the sample correlation $\rho_{k,k+1}$ over $k = 1, \ldots, p-1$ variables [4].

Although we would not expect the Toeplitz covariance estimate to be well suited to many pattern recognition applications, Hamamoto *et al.* [7] have shown, based on experiments carried out on artificial data sets, that the Toeplitz estimator can be preferable to the Van Ness [23] and orthogonal expansion [15] estimators, particularly in small-training-sample-size and high-dimensional situations.

V. New Covariance Estimate

In several image recognition applications, especially in biometric ones, the pattern classification task is commonly performed on preprocessed or well-framed patterns and the sources of variation are often the same from class to class. As a consequence, similarities in covariance shape may be assumed for all classes.

In such situations and when the sample group covariance matrices S_i are singular or not accurately estimated, linear combinations of S_i and a well-defined covariance matrix, e.g., the pooled covariance matrix S_p, may lead to a "loss of covariance information" [21]. This statement, which will be discussed first in Section V-A, forms the basis of the new covariance estimate called the maximum entropy covariance selection (MECS) method.

A. "Loss of Covariance Information"

The theoretical interpretation of the "loss of covariance information" can be described as follows. Let a matrix S_i^{mix} be given by the following linear combination:

$$S_i^{\text{mix}} = aS_i + bS_p \tag{14}$$

where the mixing parameters a and b are positive constants, and the pooled covariance matrix S_p is a nonsingular and well-defined matrix. The S_i^{mix} eigenvectors and eigenvalues are given respectively by the matrices Φ_i^{mix} and Λ_i^{mix}.

From the covariance spectral decomposition formula described in (2), it is possible to write

$$(\Phi_i^{\text{mix}})^T S_i^{\text{mix}} \Phi_i^{\text{mix}} = \Lambda_i^{\text{mix}} = \begin{bmatrix} \lambda_1^{\text{mix}} & & & 0 \\ & \lambda_2^{\text{mix}} & & \\ & & \ddots & \\ 0 & & & \lambda_p^{\text{mix}} \end{bmatrix} \tag{15}$$

where $\lambda_1^{\text{mix}}, \lambda_2^{\text{mix}}, \ldots, \lambda_p^{\text{mix}}$ are the S_i^{mix} eigenvalues and p is the dimension of the measurement space considered. Using the information provided by (14), (15) can be rewritten as

$$\begin{aligned} \Lambda_i^{\text{mix}} &= \text{diag}[\lambda_1^{\text{mix}}, \lambda_2^{\text{mix}}, \ldots, \lambda_p^{\text{mix}}] \\ &= (\Phi_i^{\text{mix}})^T \left[aS_i + bS_p\right] \Phi_i^{\text{mix}} \\ &= a(\Phi_i^{\text{mix}})^T S_i \Phi_i^{\text{mix}} + b(\Phi_i^{\text{mix}})^T S_p \Phi_i^{\text{mix}} \\ &= aZ^i + bZ^p. \end{aligned} \tag{16a}$$

The matrices Z^i and Z^p are not diagonal matrices because Φ_i^{mix} does not necessarily diagonalises both matrices simultaneously. However, as Φ_i^{mix} is the eigenvectors matrix of the linear combination of S_i and S_p, the off-diagonal elements of Z^i and Z^p necessarily cancel each other in order to generate the eigenvalues matrix Λ_i^{mix}. Therefore, the string of equalities in (16a) can be extended to

$$\begin{aligned} \Lambda_i^{\text{mix}} &= aZ^i + bZ^p \\ &= \text{diag}[a\zeta_1^i, a\zeta_2^i, \ldots, a\zeta_p^i] \\ &\quad + \text{diag}[b\zeta_1^p, b\zeta_2^p, \ldots, b\zeta_p^p] \\ &= \text{diag}[a\zeta_1^i + b\zeta_1^p, a\zeta_2^i \\ &\qquad + b\zeta_2^p, \ldots, a\zeta_p^i + b\zeta_p^p] \end{aligned} \tag{16b}$$

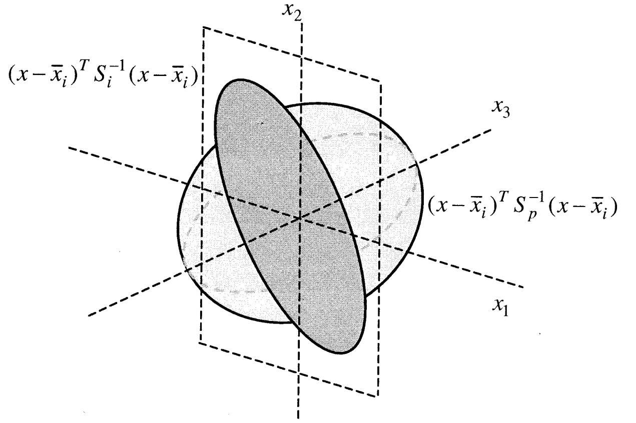
x_2
$(x-\bar{x}_i)^T S_i^{-1}(x-\bar{x}_i)$
x_3
$(x-\bar{x}_i)^T S_p^{-1}(x-\bar{x}_i)$
x_1

Fig. 1. Geometric idea of a hypothetical "loss of covariance information."

where $\zeta_1^i, \zeta_2^i, \ldots, \zeta_p^i$ and $\zeta_1^p, \zeta_2^p, \ldots, \zeta_p^p$ are, respectively, the variances of the sample and pooled covariance matrices spanned by the S_i^{mix} eigenvectors matrix Φ_i^{mix}. Then, according to (3), the inverse of S_i^{mix} becomes

$$(S_i^{\text{mix}})^{-1} = \sum_{k=1}^{p} \frac{\phi_{ik}^{\text{mix}} (\phi_{ik}^{\text{mix}})^T}{a\zeta_k^i + b\zeta_k^p} \tag{17}$$

where ϕ_{ik}^{mix} is the corresponding kth eigenvector of the matrix S_i^{mix}.

The inverse of S_i^{mix} described in (17) considers the dispersions of sample group covariance matrices spanned by all the S_i^{mix} eigenvectors. However, when the class sample sizes n_i are small or not large enough compared with the dimension of the feature space p, the corresponding lower dispersion values are often estimated to be 0 or approximately 0, implying that these values are not reliable. Therefore, a linear combination of S_i and S_p that uses the same parameters a and b as defined in (14) for the whole feature space fritters away some pooled covariance information.

The geometric idea of a hypothetical "loss of covariance information" on a three-dimensional (3-D) feature space is illustrated in Fig. 1. The constant probability density contour of S_i and S_p are represented by the 2-D (x_1, x_2) dark grey ellipse and 3-D (x_1, x_2, x_3) light grey ellipsoid, respectively. As can be seen, S_i is well defined on the plane (x_1, x_2) but not defined at all on (x_1, x_2, x_3). In fact, there is no information from S_i on the x_3 axis. As a consequence, a linear combination of S_i and S_p that shrinks or expands both matrices equally all over the feature space simply ignores this evidence. Other covariance estimators, especially the ones developed for QDF classifiers, have not addressed this problem.

B. MECS Method

The MECS method considers the issue of combining the sample group covariance matrices and the pooled covariance matrix based on the maximum entropy (ME) principle, stated briefly as

"When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have [14]."

In the problem of estimating covariance matrices for Gaussian classifiers, it is known that different covariance estimators should be optimal depending not only on the true covariance statistics of each class, but also on the number of training patterns, the dimension of the feature space, and even the ellipsoidal symmetry associated with the Gaussian distribution [3], [13]. In fact, such covariance optimization can be viewed as a problem of estimating parameters of Gaussian probability distributions under uncertainty. Therefore, the ME criterion that maximizes the uncertainty under an incomplete information context should be a promising solution.

Let a p-dimensional sample X_i be normally distributed with true mean μ_i and true covariance matrix Σ_i, i.e., $X_i \sim N_p(\mu_i, \Sigma_i)$. The entropy h of such multivariate distribution can be written as

$$h(X_i) = \frac{p}{2} + \frac{1}{2}\ln|\Sigma_i| + \frac{p}{2}\ln 2\pi \tag{18}$$

which is simply a function of the determinant of Σ_i and is invariant under any orthonormal transformation [4]. Thus, when Φ_i consists of p eigenvectors of Σ_i, we have

$$\ln\left|\Phi_i^T \Sigma_i \Phi_i\right| = \ln|\Lambda_i| = \sum_{k=1}^{p} \ln \lambda_k. \tag{19}$$

In order to maximize (19) or equivalently (18), we must select the covariance estimation of Σ_i that gives the largest eigenvalues.

Considering convex combinations between the sample group covariance S_i and S_p matrices, (19) can be rewritten [by using (16)] as

$$\ln\left|\left(\Phi_i^{\text{mix}}\right)^T (aS_i + bS_p)\,\Phi_i^{\text{mix}}\right| = \sum_{k=1}^{p} \ln(a\zeta_k^i + b\zeta_k^p) \tag{20}$$

where $\zeta_1^i, \zeta_2^i, \ldots, \zeta_p^i$ and $\zeta_1^p, \zeta_2^p, \ldots, \zeta_p^p$ are the variances of the sample and pooled covariance matrices spanned by Φ_i^{mix}, and the parameters a and b are nonnegative and sum to 1.

Moreover, as the natural logarithm is a monotonic increasing function, we do not change the problem if instead of maximizing (20) we maximize

$$\sum_{k=1}^{p} (a\zeta_k^i + b\zeta_k^p). \tag{21}$$

However, $a\zeta_k^i + b\zeta_k^p$ is a convex combination of two real numbers and the following inequality is valid [9]:

$$a\zeta_k^i + b\zeta_k^p \leq \max(\zeta_k^i, \zeta_k^p) \tag{22}$$

for any $1 \leq k \leq p$ and convex parameters a and b that are nonnegative and sum to 1. Equation (22) shows that the maximum of $a\zeta_k^i + b\zeta_k^p$ depends on k and is attained at the extreme values of the convex parameters, that is, either $a = 1$ and $b = 0$ or $a = 0$ and $b = 1$.

One possible way to maximize (21) and consequently the entropy given by the convex combination of S_i and S_p is to select the maximum variances of the sample and pooled covariance matrices given by an orthonormal projection basis that diagonalizes an unbiased $(a = b)$ linear mixture of the corresponding matrices. Recalling the assumption made that all classes have similar covariance shapes, it is reasonable to expect that the dominant eigenvectors (i.e., the eigenvectors with largest eigenvalues) of this unbiased mixture would be mostly orientated by the eigenvectors of the covariance matrix with largest eigenvalues. The choice of sample group or pooled information is then made purely on the size of the eigenvalue, which reflects the reliability of the information. Since any unbiased combination of S_i and S_p gives the same set of eigenvectors, an orthonormal basis that would avoid the loss of covariance information is the one composed of the eigenvectors of the covariance matrix given by $S_i + S_p$.

Therefore, the MECS estimator S_i^{mecs} can be calculated by the following procedure.

1) Find the eigenvectors Φ_i^{me} of the covariance given by $S_i + S_p$.

2) Calculate the variance contribution of both S_i and S_p on the Φ_i^{me} basis, i.e.,

$$\begin{aligned} \text{diag}(Z^i) &= \text{diag}[(\Phi_i^{\text{me}})^T S_i \Phi_i^{\text{me}}] = [\zeta_1^i, \zeta_2^i, \ldots, \zeta_p^i] \\ \text{diag}(Z^p) &= \text{diag}[(\Phi_i^{\text{me}})^T S_p \Phi_i^{\text{me}}] = [\zeta_1^p, \zeta_2^p, \ldots, \zeta_p^p]. \end{aligned} \tag{23}$$

3) Form a new variance matrix based on the largest values, that is,

$$Z_i^{\text{me}} = \text{diag}[\max(\zeta_1^i, \zeta_1^p), \ldots, \max(\zeta_p^i, \zeta_p^p)]. \tag{24}$$

4) Form the MECS estimator

$$S_i^{\text{mecs}} = \Phi_i^{me} Z_i^{\text{me}} (\Phi_i^{\text{me}})^T. \tag{25}$$

The MECS approach is a direct procedure that deals with the singularity and instability of S_i when similar covariance matrices are linearly combined. It does not require an optimization procedure and consequently its estimation, differently from RDA and LOOC ones, is not exclusive to the quadratic discriminant classifier. In fact, MECS can be used in the parametric quadratic classifier as well as in the nonparametric Gaussian Parzen classifier whenever the sample group covariance matrices are poorly estimated or ill posed.

VI. Experiments

In order to investigate the performance of MECS compared with the parametric QDF, LDF, RDA, and LOOC classifiers, and also with the nonparametric PZW, Van Ness, and Toeplitz classifiers, three biometric applications were considered: face recognition, facial expression recognition, and fingerprint classification.

In the face and facial expression recognition applications, the training sample sizes were chosen to be extremely small and small, respectively, compared to the dimension of the feature space. In contrast, moderate and large training sample sizes compared with the number of features were considered for the fingerprint problem.

Since in all of the applications the same number of training examples per class was considered, the prior probabilities were assumed to be equal for all classes and recognition tasks. Moreover, the RDA optimization grid was taken to be the outer product of $\lambda = [0, 0.125, 0.354, 0.65, 1.0]$ and $\gamma = [0, 0.25, 0.5, 0.75, 1.0]$, as suggested by Friedman's work [3]. Analogously, the size of the LOOC parameter was $\alpha_i = [0, 0.25, 0.5, \ldots, 2.75, 3.0]$, as given by [8], and the Van

Fig. 2. Some example images from the FERET database.

Ness smooth parameter was $\beta = [0.2, 0.4, 0.6, \ldots, 1.6, 1.8]$, as suggested by [23]. For simplicity, the Parzen window parameter h_i was assumed to be equal for all classes in all applications, and its optimum value was determined using the following set of ten values: 0.001, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 100, 1000 which are usually sufficient to empirically determine h_i^{opt} [19].

A. Face-Recognition Experiment

In the face-recognition experiments, the FERET Database [18] was used. The FERET database is the U.S. Army Face Recognition Technology facial database that has become the standard data set for benchmark studies.

Sets containing four "frontal b series" images for each of 200 total subjects were considered. Each image set is composed of a regular facial expression (referred to as "ba" images in the FERET database), an alternative expression ("bj" images), and two symmetric images ("be" and "bf" images) taken with the intention of investigating 15° pose angle effects.

For implementation convenience, all images were first resized to 96×64 pixels and transformed into eigenfeature vectors using principal component analysis (PCA) [22]. Each experiment was repeated 25 times using several of those eigenfeatures. Distinct training and test samples were randomly drawn, and the mean of the recognition rate was calculated. Since the LOOC computation requires at least three examples in each class [8], the recognition rate was computed utilizing for each subject three images to train and one image to test. Fig. 2 illustrates some example images from the FERET database cropped to the size of 96×64 pixels.

B. Facial-Expression-Recognition Experiment

Tohoku University has provided the database for the facial-expression experiment [16]. This database is composed of 193 images of expressions posed by nine Japanese females. Each person posed three or four examples of each six fundamental facial expression: anger, disgust, fear, happiness, sadness, and surprise. The database has at least 29 images for each fundamental facial expression. Fig. 3 illustrates some example images from the Tohoku Facial Expression database cropped to the size of 64×64 pixels.

Fig. 3. Some example images from the Tohoku Facial Expression database.

Analogously to the face-recognition experiments, first PCA reduces the dimensionality of the original images (which were resized to 64×64 pixels for implementation convenience) and second the discriminant Bayes' rule using the parametric and nonparametric aforementioned approaches were applied. Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean of the recognition rate was calculated. The training and test sets were respectively composed of 20 and nine images.

C. Fingerprint Classification Experiment

The fingerprint classification was performed utilizing the training and test feature vectors extracted from the grey scale images of the standard U.S. National Institute of Standards and Technology (NIST) Special Database 4 [24]. Each feature vector consists of 112 floating point numbers, made by a feature-selection procedure that ends with the PCA transform.

The fingerprints were classified into one of five categories (arch, left loop, right loop, tented arch, and whorl) with an equal number of prints from each class (400). There are 2000 first-rolling fingerprint feature vectors for training and 2000 corresponding second-rolling ones for testing. Fig. 4 illustrates some example images taken from the fingerprint database and displayed on the NIST Special Database 4 web site.[1]

VII. Results of the Parametric Classifiers

Fig. 5 presents the test average recognition error of the FERET face database. Since only three face images were used to train the classifiers, the sample group covariance matrices S_i were singular and the QDF could not be calculated. Instead, the recognition rate of the Euclidean distance classifier (EUC) that corresponds to the classical eigenfaces method proposed by Turk and Pentland [22] are displayed. Fig. 5 shows that for all the feature components considered the MECS quadratic classifier performed as well or better than the other classifiers. The MECS quadratic classifier achieved the lowest classification error—2.2%—on 50 eigenfeatures. In this application where

[1] Available. [Online]. http://www.nist.gov/srd/nistsd4.htm

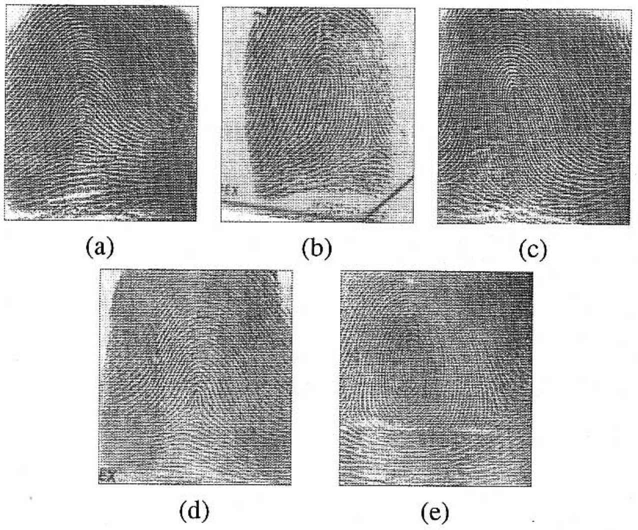

Fig. 4. (a) Arch, (b) left loop, (c) right loop, (d) tented arch, and (e) whorl.

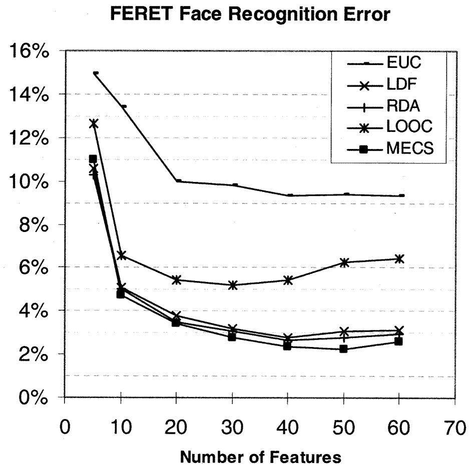

Fig. 5. FERET face database recognition error for parametric classifiers.

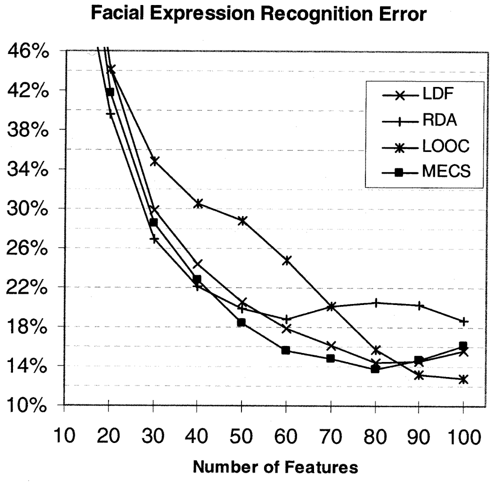

Fig. 6. Tohoku facial expression recognition error for parametric classifiers.

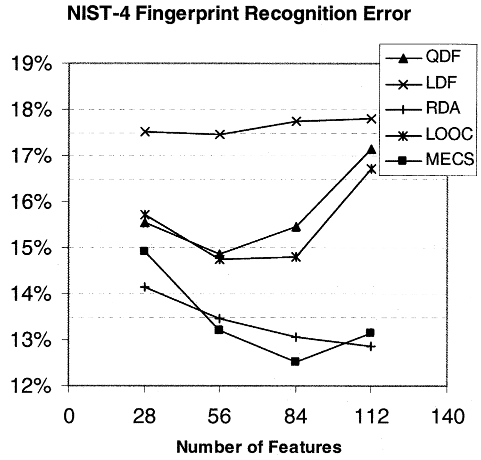

Fig. 7. NIST-4 fingerprint recognition error for parametric classifiers.

all S_i seem to be quite similar, favoring the LDF performance, the MECS classifier did better by using the full covariance information.

The results of the Tohoku facial expression recognition are presented in Fig. 6. Due to the fact that 20 images were used to form the training set of each class, the sample group covariance estimate (QDF) results were limited to 19 PCA features. These results were much less accurate than the others and were not plotted on Fig. 6. As can be seen, there is no overall dominance of any parametric covariance estimator. In lower dimension space, RDA led to lower classification error rates, followed by MECS, LDF and LOOC. When the dimensionality increased and the true covariance matrices became apparently equal and highly ellipsoidal, RDA performed poorly while MECS, LDF and LOOC improved. In the highest dimensional space the LOOC optimization, which considers the diagonal elements of the pooled estimate, took advantage of the equal-ellipsoidal behavior (for more than 70 PCAs all α_i parameters are closer to the value 3) achieving the lowest recognition error rate—12.8%—for all PCA components calculated. In this recognition application, all the computed covariance estimators were quite sensitive to the choice of the training and test sets.

The recognition results of the NIST-4 fingerprint database are presented in Fig. 7. In the lowest and highest dimension spaces (28 and 112 features), RDA led to lower classification error than MECS estimator. However, for 56 and 84 features the MECS performed better than the other classifiers. Although in this application the ratio of the training sample size to the number of features is moderate and large, favoring the QDF, RDA and LOOC classifiers, the MECS estimator achieved the lowest classification error—12.5%—on 84 components. Putting this result in perspective, a classification error of 12% but with 10% rejection of the fingerprints was reported on the same training and test sets [24].

These results confirm the findings of several researchers that choosing a nonconventional Bayesian parametric classifier between the linear and quadratic ones improves the classification accuracy in settings for which sample sizes are small and the number of features is large [3], [5], [6], [8], [20]. However, MECS new covariance approach shows that, in high-dimensional classification problems where limited training sample sizes are provided, the problem of estimating

TABLE I
COMPUTATIONAL TIME (IN SECONDS) FOR THE QUADRATIC CLASSIFIERS

Application Features	RDA	LOOC	MECS
Face			
10	1392.42	2.95	0.04
20	1860.55	7.62	0.14
30	5549.83	23.38	0.56
40	8488.75	36.08	0.98
50	10999.77	57.08	1.75
60	14644.63	73.05	2.47
Facial Expression			
10	13.02	0.73	0.01
30	20.48	3.11	0.03
50	44.99	8.37	0.06
70	87.07	17.95	0.13
90	148.73	32.29	0.24
Fingerprint			
28	247.24	38.00	0.02
56	953.39	188.47	0.04
84	2106.20	452.56	0.13
112	4251.45	934.06	0.34

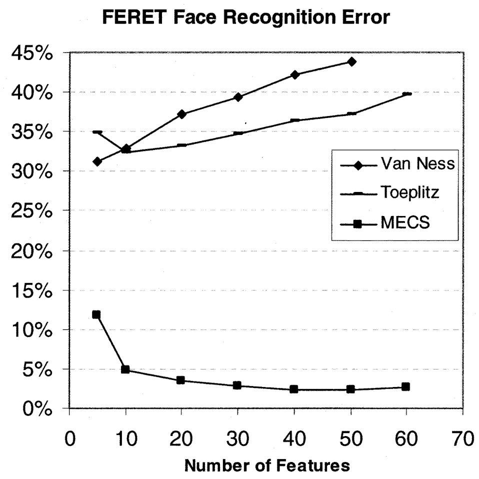

Fig. 8. FERET face database recognition error for Parzen classifiers.

covariance matrices for nonconventional quadratic classifiers is essentially an issue of combining reliable information rather than optimising classification or likelihood indexes of well-behaved samples. Table I illustrates the CPU times for our RDA, LOOC and MECS implementations on a 1-GHz desktop using a Windows-based mathematical package, given as inputs the corresponding covariance matrices S_i, S_p, I, and S.

VIII. RESULTS OF THE PARZEN WINDOW CLASSIFIERS

In this section, the results of the Gaussian Parzen Window classifiers using the sample group (PZW), Van Ness, Toeplitz, and MECS covariance estimates are analyzed.

Fig. 8 presents the test average recognition error of the FERET face database. Since only three face images were used to train the classifiers, the sample group covariance matrices were singular and the standard Gaussian Parzen Window classifier (PZW) could not be calculated. As can be seen, the MECS estimator significantly improved the face classification accuracy of the Parzen classifier compared with the other estimation approaches.

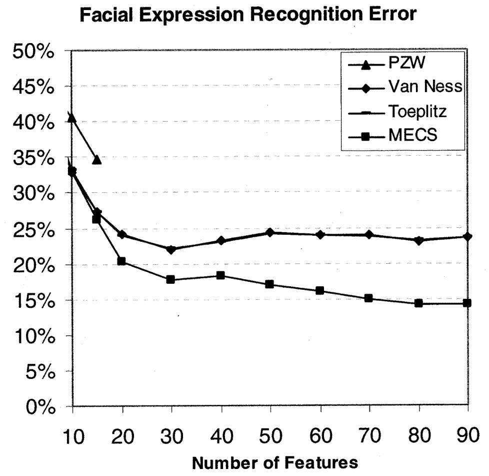

Fig. 9. Tohoku facial expression recognition error for Parzen classifiers.

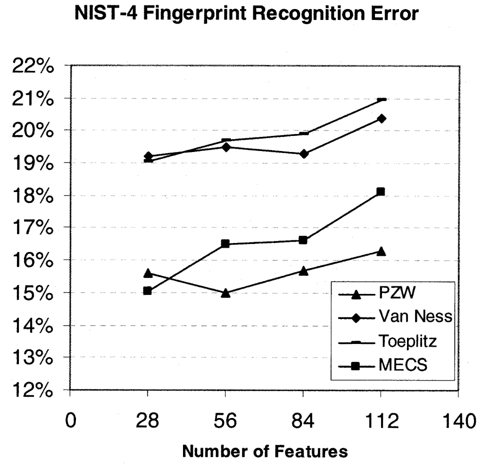

Fig. 10. NIST-4 fingerprint recognition error for Parzen classifiers.

An analogous performance of the MECS estimator is shown in Fig. 9. It presents the test average recognition error of the Tohoku facial expression database. Since 20 images were used to form the training set of each facial expression, the results relative to the standard Parzen classifier with the sample group covariance estimate (PZW) were limited. Once more, the superiority of the MECS estimator is clear regarding the Van Ness and Toeplitz estimators. In both face and facial expression applications, where the sample sizes are extremely small and small compared to the dimensionality of the feature space, the restrictive forms of the Van Ness and Toeplitz approaches underperform the potential recognition accuracy of the Gaussian Parzen classifier.

Fig. 10 presents the recognition results of the NIST-4 fingerprint database. As expected, since in this application the sample group covariance matrices are well posed estimated with 400 training patterns per class, the standard Parzen classifier (PZW) in all but one experiment led to lower recognition error than did the MECS, Van Ness, and Toeplitz estimators. According to the number of features considered, MECS achieved its best recognition error result—15.05%—when the dimensionality of the patterns was reduced to 28 components. This result was slightly worse than the PZW best result—15.00% of recognition error

TABLE II
COMPUTATIONAL TIME (IN SECONDS) FOR THE PARZEN CLASSIFIERS

Application Features	Van Ness	Toeplitz	MECS
Face			
10	943.66	0.13	0.07
20	1370.39	0.27	0.22
30	2020.80	0.39	0.51
40	2893.75	0.54	1.04
50	4016.77	0.83	1.85
60	5428.72	1.19	2.88
Facial Expression			
10	15.56	0.01	0.01
30	54.53	0.01	0.02
50	127.81	0.03	0.05
70	236.55	0.06	0.12
90	383.28	0.08	0.24
Fingerprint			
28	12754.39	0.02	0.01
56	42181.72	0.08	0.07
84	91019.83	0.22	0.17
112	161273.64	0.68	0.40

using 56 preprocessed features. Comparing the MECS classification performance with both the Van Ness and Toeplitz estimators, MECS showed again its superiority in discriminating well-framed image patterns.

The computational time for the Parzen classifiers is shown in Table II. Due to the calculation of the eigenvalues and eigenvectors, the MECS time is sometimes higher than the Toeplitz approach.

IX. CONCLUSION

In this paper, the performance of a new covariance estimate for parametric and nonparametric Bayesian classifiers was evaluated in biometric recognition problems involving small, moderate, and large training sets, large number of features, and several classes.

The new covariance estimate, called the MECS method, uses the principle of maximizing the information under an incomplete and consequently uncertainty context rather than optimizing classification accuracy or group likelihood. We believe that this is the correct approach when limited sample sizes are available in image recognition problems.

Experiments were carried out to analyze the classification performance of MECS compared with QDF, LDF, RDA, and LOOC parametric classifiers, and PZW, Van Ness, and Toeplitz nonparametric classifiers on three different well-framed image applications: face recognition, facial expression recognition, and fingerprint classification. The MECS performed at least as well as any other method and at a much lower computational cost. Since biometric applications are typically small or limited-sample-size problems, we would expect that the relative performance of the different methods investigated would not be affected by the way that image features are extracted. However, further analyses such as the false acceptance/rejection indexes for face recognition and the confusion matrix for fingerprint classification would be necessary in order to implement the new covariance estimation as a recognition system in practice.

It has been suggested (see, e.g., [23]) that, even when the underlying data are ideal for the QDF classifier, that is, the sample data of each group are unimodal with a local maximum, the nonparametric classifiers formed by nonconventional covariance estimators, such as the Van Ness approach, could achieve superior classification accuracy in limited samples and high-dimensional problems. The results of the experiments in this paper suggest that what has been behind these surprising findings is an unfair comparison between a poor covariance estimate given by the conventional maximum-likelihood approach used in the QDF classifier and a more reliable nonconventional covariance estimate used in the nonparametric classifier. When the less restricted MECS covariance estimate is used in both parametric and nonparametric classifiers, our results indicate that the parametric classifier is the best choice. Furthermore, parametric classifiers are simpler and faster to compute.

ACKNOWLEDGMENT

The authors would like to thank the referees for their constructive comments which helped to improve this paper. Also, portions of the research in this paper use the FERET database of facial images collected under the FERET program.

REFERENCES

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.

[2] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa, and C. L. Wilson, "Evaluation of pattern classifiers for fingerprint and OCR applications," *Pattern Recognit.*, vol. 27, pp. 485–501, 1994.

[3] J. H. Friedman, "Reguralized discriminant analysis," *J. Amer. Statistic. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston, MA: Academic, 1990.

[5] T. Greene and W. S. Rayens, "Partially pooled covariance matrix estimation in discriminant analysis," *Commun. Statistics-Theory Meth.*, vol. 18, no. 10, pp. 3679–3702, 1989.

[6] ——, "Covariance pooling and stabilization for classification," *Comput. Stat. Data Anal.*, vol. 11, pp. 17–42, 1991.

[7] Y. Hamamoto, Y. Fujimoto, and S. Tomitan, "On the estimation of a covariance matrix in designing Parzen classifiers," *Pattern Recognit.*, vol. 29, no. 10, pp. 1751–1759, 1996.

[8] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 763–767, July 1996.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[10] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1982, vol. 2, pp. 835–855.

[11] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4–37, Jan. 2000.

[12] A. K. Jain and M. D. Ramaswami, "Classifier design with parzen windows," in *Pattern Recognition and Artificial Intelligence*, E. S. Gelsema and L. N. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1988, pp. 211–228.

[13] M. James, *Classification Algorithms*. London, U.K.: William Collins, 1985.

[14] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, Sept., 1982.

[15] F. Kimura, K. Takashima, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to chinese character recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 149–153, Jan. 1987.

[16] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1357–1362, Dec. 1999.

[17] S. Omachi, F. Sun, and H. Aso, “A new approximation method of the quadratic discriminant function,” in *Proc. SSPR&SPR 2000*, vol. LNCS 1876, 2000, pp. 601–610.

[18] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, “The FERET database and evaluation procedure for face recognition algorithms,” *Image Vis. Comput. J.*, vol. 16, no. 5, pp. 295–306, 1998.

[19] S. J. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 252–264, Mar. 1991.

[20] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, “Using mixture covariance matrices to improve face and facial expression recognitions,” in *Proc. 3rd Int. Conf. Audio- and Video-Based Biometric Person Authentication*, vol. LNCS 2091, Halmstad, Sweden, June 2001, pp. 71–77.

[21] ——, “A new quadratic classifier applied to biometric recognition,” in *Proc. Post-ECCV Int. Workshop Biometric Authentication*, vol. LNCS 2359, Copenhagen, Denmark, June 2002, pp. 186–196.

[22] M. Turk and A. Pentland, “Eigenfaces for recognition,” *J. Cognitive Neurosci.*, vol. 3, pp. 71–86, 1991.

[23] J. Van Ness, “On the dominance of nonparametric bayes rule discriminant algorithms in high dimensions,” *Pattern Recognit.*, vol. 12, pp. 355–368, 1980.

[24] C. L. Wilson, G. T. Candela, P. J. Grother, C. I. Watson, and R. A. Wilkinson, “Massively parallel neural network fingerprint classification system,” National Inst. of Standards and Technology, Tech. Rep. NIST IR 4880, 1992.

Carlos E. Thomaz received the B.Sc. degree in electronic engineering and the M.Sc. degree in electrical engineering from the Catholic University of Rio de Janeiro (PUC-RJ), Rio de Janeiro, Brazil, in 1993 and in 1999, respectively, and he is currently working toward the Ph.D. degree in biometrics recognition at Imperial College London, London, U.K.

In October 2000, he joined the Department of Computing, Imperial College London. His general interests are in computer vision, statistical pattern recognition, and machine learning, while his specific research interests are in limited-sample-size problems in pattern recognition.

Duncan F. Gillies received the degree in engineering science from Cambridge University, Cambridge, U.K., in 1971, the M.Sc. degree in computing from Birkbeck College London, London, U.K., and the Ph.D. degree in the area of artificial intelligence from Queen Mary College, London, U.K.

After teaching for six years at the Polytechnic of the South Bank, he moved to the Department of Computing, Imperial College London, London, U.K., in October 1983 where he is now a Reader. He has worked in the areas of computer graphics and vision, and their applications in the medical field, and in probabilistic inference methods.

Raul Q. Feitosa received the B.Sc. degree in electronic engineering and the M.Sc. degree in engineering from the Technological Aeronautics Institute (ITA), São José dos Campos, Brazil, in 1979 and 1983, respectively, and the Dr.-Ing. degree in computer architecture from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1988.

After working for five years in industry, he started his doctorate research with the University of Erlangen-Nürnberg. Since then he has worked with the Department of Electrical Engineering, Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, and with the Department of Computer Engineering, State University of Rio de Janeiro, Rio de Janeiro, Brazil, where he is an Associate Professor. His research interests include vision, pattern recognition, and their applications in biometrics and interpretation of remotely sensed images.