Covariance matrix estimation for appearance-based face image processing

John A Robinson
Department of Electronics, University of York, YO10 5DD
jar11@ohm.york.ac.uk

Abstract

In appearance-based image processing, high-dimensional statistical models are estimated from low numbers of training samples. Sample scatter matrices are unreliable estimators of class covariances, yet many methods rely on them for dimensionality reduction and often for classification too. This paper argues for regularized covariance estimation and introduces a new method suitable for appearance-based image processing. The method is demonstrated for face detection, where a maximum likelihood classifier trained with regularized covariances achieves discrimination and detection results comparable to those of complicated multimodal and non-linear classifiers.

1 Introduction

Appearance-based approaches to object detection and recognition treat all n pixels of an image equally, stacking them into a measurement vector which is analysed according to models derived from the statistics of training samples. They avoid assumptions about geometric or structural image features, depending instead on measured dependencies in the training set.

Since the earliest work in appearance-based analysis (e.g. [1]), it has been acknowledged that the high dimensionality of image space is a problem. Often the number of training images available N is less than the dimensionality n and therefore much less than M, the number required to estimate a distribution's parameters directly. M's value depends on the model assumed for the multivariate distribution, but it is certainly much more than n. Dimensionality reduction is sometimes offered as a solution to this problem, but must itself rely on statistics. Almost always the total scatter matrix is used as an estimate for a global covariance matrix and a process then selects combinations of measurements that are supposed to characterize the class. Usually the transform which maps into the subspace also scales along the new feature axes. Principal Components Analysis (often used for dimensionality reduction even when more sophisticated classifiers are used within the subspace) derives the transform and the scalings simultaneously as the eigenvectors and eigenvalues of the estimated covariance matrix. The problem for PCA of $N < n$ is soluble by a simple algebraic trick [2]. But this does nothing for the real problem, which is that $N << M$ and that the scatter matrix is therefore certainly not a reliable estimate for the covariance matrix.

Dimensionality reduction is so common as a first stage that appearance-based classification is almost synonymous with subspace methods. For face processing, for example, Shakhnarovich and Moghaddam begin their review chapter [3] (titled "Face Recognition in Subspaces") with the statement that faces are within a submanifold of low intrinsic dimensionality, while Penev and Sirovich [4] seek to discover this intrinsic dimensionality by experiment. Yet the only sustainable reason for reduction to a subspace is to allow more efficient computation. The suggestion that a particular face's non-zero component outside face space is "due to sensor noise" [3] mistakes an approximation (allowing all face variability in a low-variance dimension to be approximated by zero) for an essential property. Singular or almost-singular scatter matrices certainly do not provide reasons to suppose that face space has anything less than full dimensionality: there is little chance that they adequately approximate the class covariance.

BMVC 2005 doi:10.5244/C.19.43

The estimation of covariance matrices is not just for dimensionality reduction. In subsequent class modelling, the sample to prototype distances may be adjusted according to class covariances, or, if a multimodal model is assumed, the subclass covariances. Again an eigenspace representation may be used, though more recent developments in appearance-based processing have tended to use non-linear classifiers or multiple kernels [5,6]. Even so, the reliability of statistical estimates of probability distributions remains a central question.

Section 2 of this paper reviews options for improving estimates from scatter matrices to get more reliable covariances. Section 3 then introduces a new method of covariance matrix estimation particularly suited to appearance-based processing. Section 4 reports experiments on model selection – that is, the training of mixing parameters in the new method.

To the author's knowledge, no previous paper has reported performance for one of the most fundamental of statistical classifiers – the maximum likelihood classifier for unimodal normal distributions – for tasks like face detection. Perhaps the reason for this is that ML classification simply does not work at all when the covariance matrices for the class densities are incorrectly estimated. With good estimations, ML classification provides a viable baseline for face detection (for example), as section 5 shows.

2 Covariance matrix estimation

2.1 Notation and assumptions

Suppose that each of K classes is characterized, on the basis of n-dimensional training samples, by a multivariate normal distribution:

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}\left|\mathbf{\Sigma}_i\right|^{-1/2}\exp[-1/2(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] \qquad (1)$$

where $\boldsymbol{\mu}_i$ is the class's mean vector, $\mathbf{\Sigma}_i$ is its covariance matrix, and T denotes transpose. While not all that follows depends on the normality assumption, the final classification tasks and therefore the optimization criterion use the Maximum Likelihood classifier which assumes normality, namely, classify a sample $\mathbf{x}$ as belonging to class k if

$$d_k(\mathbf{x}) = \min_{1\le i\le K} d_i(\mathbf{x}) \qquad (2)$$

with

$$d_i(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln\left|\mathbf{\Sigma}_i\right| \qquad (3)$$

which, if the prior probabilities P_i of the classes are known, is related to the Bayes classifier

$$d_i(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln\left|\mathbf{\Sigma}_i\right| - 2\ln P_i \qquad (4)$$

by a scalar offset.

The use of a unimodal multivariate normal model may be inappropriate for appearance-based image processing because it assumes a compact, convex distribution with ellipsoidal symmetry. (The particular shape of the normal distribution relative to other compact, convex, symmetric distributions, has little effect for classification [7].) For example, if the images shown in figure 1(a) and 1(b) are to be included in a particular class, then either 1(c) is also in the class or the class is not convex.

(a) (b) (c)

Figure 1. Non-convexity of a possible face class

The question of identifying multi-modality and concavity is not addressed here, so the problem illustrated by figure 1 remains. However, section 5 will show that in practice a unimodal normal model is effective for a real classification task.

With the normal model, the characterization of class *i* amounts to estimating the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The estimated values are denoted $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$. For $\hat{\boldsymbol{\mu}}_i$ the mean of the class *i* training samples is the maximum likelihood estimate, and there is no reason to modify this on the basis of any out-of-class samples. The class *i* training samples also provide a scatter matrix

$$\mathbf{S}_i = \tfrac{1}{N_i}\sum_{j=1}^{N_i}(\mathbf{x}_{i,j} - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_{i,j} - \hat{\boldsymbol{\mu}}_i)^T \tag{5}$$

where N_i is the number of training samples for class *i*, of which $\mathbf{x}_{i,j}$ is the *j*th.

The average scatter matrix is

$$\mathbf{S}_{average} = \tfrac{1}{K}\sum_{j=1}^{K}\mathbf{S}_j \tag{6}$$

which weights all classes equally, whereas the pooled scatter matrix is

$$\mathbf{S}_{pooled} = \tfrac{1}{N}\sum_{j=1}^{K}N_j\mathbf{S}_j \tag{7}$$

which weights all training samples equally. Finally, in some contexts, there may be a total scatter matrix, constructed from all the training samples including some which are unlabelled but belong to a superclass of which all *K* classes are part:

$$\mathbf{S}_{total} = \tfrac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T \tag{8}$$

with *N* the total number of samples and $\hat{\boldsymbol{\mu}}$ the mean vector of all samples.

2.2 Prior work

Important cases of prior art are Friedman's *Regularized Discriminant Analysis* [8] and the series of *Leave-One-Out-Covariance Matrix* estimators proposed by Landgrebe *et al* [9-12]. Although both drew on significant earlier work (for example as reviewed in [13,14]), they are now the most-used methods for small-sample high-dimensional covariance matrix estimation.

In developing *Regularized Discriminant Analysis (RDA)* [8], Friedman noted that the estimate of a class covariance matrix by its training sample scatter matrix is incorrect for any reasonable Bayesian classification rule [15]. In consequence, linear discriminant classification, where all classes are assumed have the same covariance estimated from the pooled scatter, is often superior to quadratic classification by use of the individual scatter matrices. Moreover, if there are too few samples overall to form a good estimate of the pooled covariance, it can be even better to use Euclidean distance (nearest mean) classification. Friedman therefore proposed a regularization scheme to mix each class's scatter matrix with the pooled matrix and with the identity matrix as follows.

$$\hat{\boldsymbol{\Sigma}}_i(\lambda,\gamma) = (1-\gamma)\hat{\boldsymbol{\Sigma}}_i(\lambda) + \gamma\left(\frac{tr(\hat{\boldsymbol{\Sigma}}_i(\lambda))}{n}\right)I \tag{8}$$

where

$$\hat{\boldsymbol{\Sigma}}_i(\lambda) = \frac{(1-\lambda)(N_i-1)\mathbf{S}_k + \lambda(N-K)\mathbf{S}_{pooled}}{(1-\lambda)N_i + \lambda N} \tag{9}$$

(This is actually a simplification of Friedman's original definition which allowed each training sample to be differently weighted.)

The regularization parameters λ, γ range from 0 to 1. They are estimated by minimizing the leave-one-out cross-validation errors over a grid of values, typically (0, .125, .354, .650, 1) for λ and (0, .25, .5, .75, 1) for γ. That is, the parameters are set at one of the 25 combinations of λ, γ values, then for each class, one training sample is removed in turn, the scatter matrix is estimated from the remainder, the class covariance matrix is estimated according to (8) and (9) and the missing sample is then classified. This is repeated over all samples and all classes, yielding a total number of classification errors. The λ, γ that give the lowest such total are those chosen.

Friedman reported that RDA outperformed quadratic ML classification using the individual class scatter matrices and linear ML classification using the pooled scatter matrix for a range of distributions. Effectively it adapted between quadratic and linear classification on the basis of the data. Only in the case of equal, highly ellipsoidal covariance matrices with mean differences in low-dimensional subspace was linear classification with the pooled scatter matrix able to beat RDA's performance marginally, and then only in relatively low dimensions.

Landgrebe and coauthors developed a series of estimators that sometimes outperform RDA in their tests. The main difference from RDA is that instead of using leave-one-out classification to find the optimal parameter values, they measure the log likelihoods of the left-out training samples for the class to which they belong. They can therefore optimize each class in isolation. This allows them to use different parameters for each class, and also saves significantly in computation time over RDA.

The most general estimator proposed by Landgrabe and his collaborators is Mixed-LOOC1 given by:

$$\hat{\boldsymbol{\Sigma}}_i(a_i, b_i, c_i, d_i, e_i, f_i) = a_i \frac{tr(\mathbf{S}_i)}{n}\mathbf{I} + b_i diag(\mathbf{S}_i) + c_i \mathbf{S}_i + d_i \frac{tr(\mathbf{S}_{pooled})}{n}\mathbf{I} + e_i diag(\mathbf{S}_{pooled}) + f_i \mathbf{S}_{pooled} \quad (10)$$

where $a_i + b_i + c_i + d_i + e_i + f_i = 1$ and $i = 1,2,\ldots K$

Mixed-LOOC1 subsumes Landgrebe's earlier models by including terms that weight the individual and pooled scatter matrices, their diagonals, and the identity matrix. For each class it is required to optimize over the six parameters $a_i, b_i, c_i, d_i, e_i, f_i$, and this is done by exhaustive calculation over a coarse grid, as in RDA. Noting from earlier work that the multidimensional optima usually only include a pair of non-zero parameter values, Landgrebe et al also propose a simplified version, Mixed-LOOC2, that mixes terms of equation (10) in a limited set of unequal pairs then selects the best-performing combination. For example, Mixed-LOOC2 might yield an estimator that combines the class scatter matrix and the diagonal of the pooled scatter matrix.

3 A new regularized covariance estimator (*RCE*)

3.1 Components of the estimator and their weight parameters

In this section a new covariance matrix estimator suited to appearance-based image processing is introduced. In order to motivate the design, and explain why it differs from previous approaches, it is useful to consider how the class scatter matrices may justifiably be modified by samples from outside the class.

To begin, suppose that the class scatter matrices are well-conditioned and therefore invertible with non-zero eigenvalues. In practice they will be singular because there are insufficient training samples, but one can imagine adding the smallest amount necessary for numerically-stable inversion to the eigenvalues. As a result each class has a highly ellipsoidal distribution. The important question is, what sort of operations on the hyperellipsoids make sense for regularizing the covariance matrix?

An *n*-dimensional hyperellipsoid has three fundamental properties – its volume (a scalar), its shape (the lengths of its *n* axes; *n-1* free if the volume is known) and its orientation (specified by *n-1* angles). In principle these can be manipulated independently. Changing the volume of the scatter distribution corresponds to changing the matrix determinant or scaling all the eigenvalues equally. Changing the shape corresponds to a transformation of the eigenvalues that is not a pure multiplication. Changing the orientation corresponds to multiplication by an orthogonal matrix. The manipulations actually done to the scatter matrix hyperellipsoid by RDA and Mixed-LOOC alter the three properties in combinations as follows:

1. Adding a multiple of the identity matrix (as done in both RDA and Mixed-LOOC) causes a shape change – long axes get shorter, short ones get longer – and also a volume change, despite the weights-sum-to-1 constraint.
    a. The shape change is favourable for regularization – it diminishes the importance of distances in the low-variance subspace whose eigenvalues are most dramatically affected by the scarcity of training data. The weighted sum with the identity matrix has two other valuable properties: it is the simplest mechanism for adjusting shape and it has a direct interpretation in the pixel domain as the addition of uncorrelated noise.
    b. However, the volume change is less well motivated and the use of the average of the diagonal of a scatter matrix in both RDA and Mixed-LOOC to "normalize" the identity matrix term is questionable, first because it takes the volume of the non-regularized scatter as a reliable estimator of covariance volume, and second because it corresponds in the pixel domain to adding different amounts of uncorrelated noise to each class.
2. Forming a weighted sum of the class scatter and a pooled or average scatter changes all three of volume, shape and orientation. However, it is the only justifiable way of effecting an orientation change: there are no other training data that could rotate the scatter matrix's axes. In principle, the pooled matrix could be decomposed into an re-orientation and a scaling and these applied independently, but the simple addition of matrices has the virtue of simplicity and two obvious endpoints – the class scatter and the common scatter.
3. Forming a weighted sum with the diagonal either of the class scatter or the pooled scatter (as done in Mixed-LOOC but not RDA) lacks theoretical justification. If the scales of the different dimensions are different, then there is perhaps a rationale, but it is notable that the only reported experiments in which the original LOOC significantly outperformed RDA [9] were those where the covariance eigenvectors were aligned with the measurement axes and where, therefore, the diagonal entries (the variances) would accidentally estimate the full covariance.

In the light of points 1a and 2, the new estimator uses the identity matrix and the common sample scatter as components for weighted corrections to the class sample scatter. Following point 3, it does not include matrix diagonals. The argument of 1b relates to the weighting of the components. A broader issue is whether global or class-wise weight parameters should be used.

Friedman's RDA has two global parameters which are used in the estimation of every class's covariance. In contrast, Mixed-LOOC1 has six parameters *per class*, though there is also a simplified approximation with just two per class. The new estimator adopts global parameters for combining a class's scatter matrix with the common sample scatter and the identity matrix, but recognizing that this produces different volume changes for each class, compensates by classification threshold shift parameters between each pair of classes. Section 3.3 gives details.

3.2 Objective functions and optimization method

The two global parameters of Friedman's RDA are optimized by exhaustive testing over a grid of possible values. The objective function is classification error and so all combinations of classes are tested and optimized together. Mixed-LOOC's per-class parameters are also optimized by exhaustive testing over grids of possible values. The objective function is average log likelihood of class membership which more directly measures change in ellipsoid shape

than RDA, although there is no clear reason why one should use average rather than a minimax or some other biased measure.

Appearance-based image processing uses estimated distributions for more than just classification. Dimensionality reduction and recovery of missing data are two examples. An objective function that matches the analysis task to be done may reasonably be expected to be superior to one that describes a class in isolation. For example, for face recognition, it would be appropriate to use separability of identities as a criterion in the estimation of face space (a reverse of the process in [16]). Where conditional densities are used for recovering missing data, an appropriate objective function is the mean square error between recovered and actual values in training data [17]. Similarly, the formulation that follows, which is directed towards classification, follows RDA in adopting a classification error criterion.

Testing parameters on a sparse grid is a global but coarse search for the best combination of values. If the objective function has a single optimum, a more accurate estimate for the same computation time would be to use a standard iterative multidimensional optimization scheme. It is therefore an empirical question which approach the new estimator should use. Section 4 discusses this further in the light of experimental results.

3.3 *Regularized Covariance Estimator (RCE)* Method

RCE estimates the covariance matrix for class *i* as:

$$\hat{\boldsymbol{\Sigma}}_i(\alpha,\beta) = \alpha\mathbf{S}_i + (1-\alpha)\mathbf{S}_{total} + \beta\mathbf{I} \qquad (11)$$

α goes from 0 to 1 but β is allowed to vary freely so the estimate clearly has a non-normalized effect on covariance volume (determinant). This does not affect the relative values of the eigenvalues nor the direction of the eigenvectors, so it is of no consequence in dimensionality reduction or data recovery. However it matters for classification. RCE therefore modifies the classification rule given by equation (2) to

$$d_k(\mathbf{x}) = \min(d_1(\mathbf{x}), \min_{2\le i\le K}(d_i(\mathbf{x}) + \gamma_i)) \qquad (12)$$

where the γ_i are "volume correction" scalars meant to undo the differential effects on the volume of the different scatter matrices caused by the addition of the same $\beta\mathbf{I}$ to each class in equation (11). All of these are relative to class 1 for which there is no γ_1.

α and β are tested either on a grid or at points chosen sequentially by an optimizer. At each such α,β a subset of the training samples is used to develop scatter matrices while the remainder are classified according to the covariances estimated from (11) and the ordinary ML decision rules (2), (3). A multivariate optimization is then done over the γ_i s to choose the values that when added to the ML distances as shifts will give optimal classification. This optimization is *K-1*-dimensional, so in the case of two classes, it reduces to finding one shift which minimizes the classification error rate. The shift then has a similar effect on the final classification as the determinants and *a priori* probability terms in (3) and (4) have.

4 Experiments on model selection

To examine the performance of the new covariance estimator and compare it with alternative approaches, a series of experiments on two face-image classification problems were conducted. Both of these used 19x19 greyscale pictures for a total dimensionality of 361. All training images were normalized to the same luminance mean and variance and the face training images were centred just above the nose tip. The applications were discrimination of faces from non-faces and discrimination between smiling and neutral faces. The latter is an example of a relatively small-sample-size problem. Table 1 summarizes the experiments, with each major row representing a series of related tests. For each major row, the following were done:

1. two experiments where α and β were applied over a dense grid, the optimal value of γ calculated for each case (note that there is only one γ for two-class problems), the best α, β, γ combination determined and then applied to a test set. The table shows the sizes of (a) the training subset used to estimate the scatter matrices, (b) the training subset classified by regularized estimates defined by equation (11) from which optimal values for α, β, γ were obtained, and (c) the test set classified by the final regularized classifier, for one of those two experiments. In the second experiment, the roles of the second training subset and the test set were reversed. Table 1's results are for the worse of those two cases.
2. unregularized quadratic ML classification and pooled-scatter linear ML classification experiments. The results are included in the table for comparison.
3. a pair of 1D experiments, whose detailed results are not shown. In one of these, mixtures were supplemented with weighted diagonals of the scatter matrices to provide a comparison with Mixed-LOOC-type classification, and in the second the equal weighting of the identity matrix for all classes was replaced by an average-variance weighting as used in RDA. In each case the form used in the new RCE outperformed the alternatives.

Briefly, table 1 shows that RCE provides much better performance than classification without regularization, even where the per-class sample size is over five times the dimensionality (as in the face/non-face experiment). Regularization with both the identity matrix and a pooled or common scatter matrix is beneficial, and RCE mixes these according to sample size and the nature of the data. Mixing with total scatter over many unclassified images rather than pooled scatter is only of benefit at lower sample sizes. The mixtures resulting from different sets of training samples are closely consistent.

Figure 2 shows a typical α, β optimization surface where at each point the optimal γ has been used to calculate the log error rate. In common with other optimization surfaces measured during the experiments, this has multiple minima, but a large area over which the results are close to the optimal. Rather than the iterative optimizer approach suggested in sections 3.2 and 3.3 therefore, it appears that global optimization over a fairly coarse grid will be most computationally efficient for approximate solutions. Notably, a good estimation is available by regularizing with the identity matrix alone. Since this scales but does not rotate the axes of the distribution, it suggests that PCA using the scatter matrix is appropriate for selecting highest-variance dimensions, but not for scaling within the subspace.

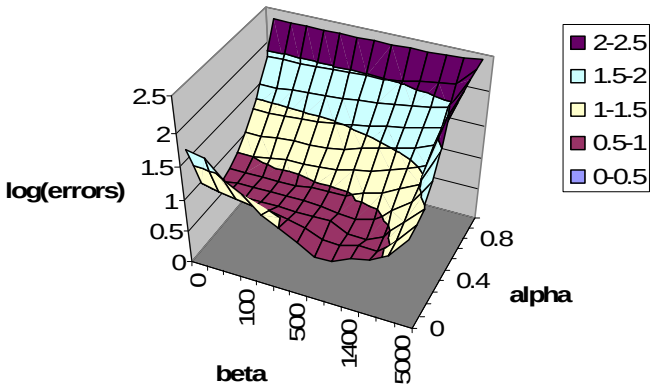

Figure 2. An optimization surface corresponding to the first major row in table 1.

5 Face Detection Application

The experiments in section 4 yield regularized face and non-face covariance matrices, so it is reasonable to apply these to face detection. Two of the best known appearance-based face

Experiment	Classes	Sample sets	Num samp-les per set	Estimated parameter values			Classif -ication errors	Error rate
				α	β	γ		
Face/non-face discrimination with large pool	Faces	Train S_i	2249	0.7	100	-46	8 out of 2646 (=1370 +1276)	0.30%
		Train α,β,γ	1369					
		Test	1370					
	Non-faces	Train S_i	3203					
		Train γ	1276					
		Test	1276					
	Unclassified	Train S_{total}	13200					
	No regularization						73	2.76%
	Pooled covariance						288	10.88%
Face/non-face discrimination with no extra pool	Faces	Train S_i	2249	1	1000	-29	4 out of 2646 (=1370 +1276)	0.15%
		Train γ	1369					
		Test	1370					
	Non-faces	Train S_i	3203					
		Train γ	1276					
		Test	1276					
	No regularization						73	2.76%
	Pooled covariance						88	3.33%
Smile/neutral discrimination with large pool	Smiles	Train S_i	120	0.5	5000	1	16 out of 205	7.80%
		Train γ	44					
		Test	69					
	Neutral	Train S_i	200					
		Train γ	100					
		Test	136					
	Unclassified faces	Train S_{total}	2249					
	No regularization						68	33.2%
	Pooled covariance						31	15.1%
Smile/neutral discrimination with no extra pool	Smiles	Train S_i	120	0.5	5000	1	14 out of 205	6.83%
		Train γ	44					
		Test	69					
	Neutral	Train S_i	200					
		Train γ	100					
		Test	136					
	No regularization						68	33.2%
	Pooled covariance						40	19.5%

Table 1. Model Selection Experiments and Results (see section 4)

detectors are due to Rowley *et al* [18] and Sung and Poggio [19]. The former uses multiple neural networks that scan a 20x20 window over scaled versions of the image while the latter uses elliptic k-means training of a multimodal structure of 6 face clusters and 6 non-face clusters with a similar scanning mechanism on 19x19 windows. Figure 3 shows examples of a face detection scanner equivalent to those of [18,19] but using maximum likelihood classification of each 19x19 window as face or non-face, according to regularized covariance matrices. The misdetection rates are comparable to those of [18] and [19]. Errors are illustrated on the bottom row of figure 3 – false positives on the left and a false negative on the right. The latter was one of only three missed faces on a much larger test set.

Having found faces as in figure 3 it is a simple matter to feed detected windows to the smiling/neutral classifier also developed in section 4. The result for Mona Lisa, for example, is neutral, though that face is particularly close to the classification boundary.

6 Conclusion

The new covariance estimator *RCE* introduced in this paper is especially suited to appearance-based image analysis. The new estimator differs from previous mechanisms by: (a)

regularizing all classes in the same way, then compensating for volume distortions via shift parameters; (b) using an application-dependent cost function.

Although just one demonstration of regularized covariance estimation, figure 3 is remarkable, not because it represents a new sophisticated classifier, but because it shows how one of the most fundamental of statistical classifiers can achieve competitive performance when the class covariance matrices are properly estimated. This suggests that maximum likelihood classification between unimodal, compact, convex, normal distributions should be the baseline against which more complicated approaches are compared, and that appearance-based image processing in general might benefit from regularized covariance estimation.

References

[1] M Kirby, L Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", *Transactions on PAMI*, Vol 12, No 1, January 1990, pp 103-108.

[2] M Turk, A Pentland, "Face Recognition Using Eigenfaces", *Proceedings of IEEE CVPR*, Maui, Hawaii, December 1991, pp 586-590.

[3] G Shakhnarovich, B Moghaddam, "Face Recognition in Subspaces", in *Handbook of Face Recognition*, eds. S Z Li, A K Jain, Springer-Verlag, 2004.

[4] P S Penev, L Sirovich, "The Global Dimensionality of Face Space", *Fourth IEEE Internat Conf Automatic Face and Gesture Recognition*, Grenoble, France, March 26-30 2000.

[5] B Moghaddam, A Pentland, "Probabilistic Visual Learning for Object Representation", *IEEE Transactions on PAMI*, Vol 19 No 7, July 1997, pp 696-710.

[6] B Schoelkopt, A Smola, K-R Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation*, Vol 10 No 5, 1998, pp1299-1319.

[7] M James, *Classification Algorithms*, Collins, London, 1985.

[8] J H Friedman, "Regularized Discriminant Analysis", *Journal of the American Statistical Association*, Vol 84 No 405, March 1989, pp 165-175.

[9] J P Hoffbeck, D A Landgrebe, "Covariance Matrix Estimation and Classification With Limited Training Images", *IEEE Transactions on PAMI*, Vol 18 No 7, July 1996, pp 763-765.

[10] L Jimenez, D A Landgrebe, "Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate Data", *IEEE Transactions on Systems, Man and Cybernetics*, Vol 28, Part C, No 1, Feb 1998, pp 39-54.

[11] S Tadjudin, D A Landgrebe, "Covariance Estimation With Limited Training Samples", *IEEE Transactions on Geoscience and Remote Sensing*, Vol 37, No 4, July 1999.

[12] B-C Kuo, D A Landgrebe, "A Covariance Estimator For Small Sample Size Classification Problems and its Application to Feature Extraction", *IEEE Transactions on Geoscience and Remote Sensing*, Vol 40, No 4, April 2002, pp 814-819.

[13] D M Titterington, "Common Structure of Smoothing Techniques in Statistics", *International Statistical Review*, Vol 53, 1985, pp 141-170.

[14] F O'Sullivan, "A Statistical Perspective on Ill-Posed Inverse Problems", *Statistical Science*, Vol 1, 1986, pp 502-527.

[15] P Enis, S Geisser, "Optimal Predictive Linear Discriminants", *The Annals of Statistics*, Vol 2, 1974, pp 403-410.

[16] V Belhumeur, J Hespanha, D Kriefman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection", *IEEE Transactions on PAMI*, Vol 19 No 7, July 1997, pp 711-720.

[17] J A Robinson, J R Hyde, "Estimation of Face Depths by Conditional Densities", *British Machine Vision Conference 2005*, Oxford, UK, September 2005.

[18] H A Rowley, S Baluja, T Kanada, "Neural Network-Based Face Detection", *IEEE Transactions on PAMI*, Vol 20 No 1, January 1998, pp 23-38.

[19] K-K Sung, T Poggio, "Example-Based Learning for View-Based Human Face Detection", *IEEE Transactions on PAMI*, Vol 20 No 1, January 1998, pp 39-51.

Figure 3. Example outputs of the Maximum Likelihood Face Detector