Appears in 2^{nd} Canadian Conference on Computer and Robot Vision, Victoria, Canada, 2005.

Minimum Bayes Error Features for Visual Recognition by Sequential Feature Selection and Extraction

Gustavo Carneiro

Department of Computer Science
University of British Columbia
carneiro@cs.ubc.ca

Nuno Vasconcelos

Department of Electrical and Computer engineering
University of California San Diego
nuno@ece.ucsd.edu

Abstract

The extraction of optimal features, in a classification sense, is still quite challenging in the context of large-scale classification problems (such as visual recognition), involving a large number of classes and significant amounts of training data per class. We present an optimal, in the minimum Bayes error sense, algorithm for feature design that combines the most appealing properties of the two strategies that are currently dominant: feature extraction (FE) and feature selection (FS). The new algorithm proceeds by interleaving pairs of FS and FE steps, which amount to a sequential search for the most discriminant directions in a collection of two dimensional subspaces. It combines the fast convergence rate of FS with the ability of FE to uncover optimal features that are not part of the original basis functions, leading to solutions that are better than those achievable by either FE or FS alone, in a small number of iterations. Because the basic iteration has very low complexity, the new algorithm is scalable in the number of classes of the recognition problem, a property that is currently only available for feature extraction methods that are either sub-optimal or optimal under restrictive assumptions that do not hold for generic recognition. Experimental results show significant improvements over these methods, either through much greater robustness to local minima or by achieving significantly faster convergence.

1. Introduction

The formulation of visual recognition as a problem of statistical classification (see e.g. [13, 2, 4, 8, 10, 15]) has resulted in solutions of unprecedented success for problems such as face detection [11, 14]. This success is due, in large part, to the fact that the statistical formulation supports feature extraction strategies that are data-driven and explicitly minimize classification error. While there are multiple ways to achieve this goal, e.g. through the search for the optimal weight configuration for the hidden nodes of a neural network [4, 11], the selection of a best set of basis functions from a predefined set [8, 14], or the selection of feature configurations [10, 15], the end product is invariably a set of features that are optimal, in the classification sense, for the recognition problem.

Typically, this is accomplished by posing the problem of feature extraction or selection as one of learning a discriminant classifier, e.g. a neural net, a support vector machine, or a boosted perceptron. While embedding feature design in the design of the overall classifier has the advantage of explicitly optimizing performance for the recognition task at hand, it also has significant drawbacks. The most challenging among these is a significant computational complexity: assuming that the initial pool of features is large, the complex problem of designing a complete classifier on a high-dimensional feature space has to be solved at each step of feature extraction. Since most of the state-of-the-art algorithms for the design of discriminant classifiers (e.g. backpropagation, SVM learning, or boosting) do not scale well with the number of classes that need to be discriminated, the task is virtually impossible in the context of large-scale recognition systems, i.e. recognition systems applicable to problems containing thousands of classes and significant amounts of training data per class. For this reason, sub-optimal feature extraction techniques such as principal component analysis (PCA) [13], or linear discriminant analysis (LDA) [2], remain the most popular for problems such as face, object, or texture recognition.

It should be noted that the simultaneous design of classifier and feature set is not a necessary requirement for achieving optimal (in a classification sense) features. In particular, it is known from Bayesian decision theory that 1) the probability of error of any classifier is lower bounded by the *Bayes error* (BE), 2) the BE only depends on the feature space, not the classifier itself, and 3) there is always at least one classifier that achieves this lower bound (the Bayes classifier). Hence, at least from a theoretical point of view, it should be possible to find the optimal feature space for a given classification problem without designing the classifier itself: it suf-

fices to find the feature transformation that leads to the feature space where the BE is minimum. By avoiding the complexity of iterated classifier design, this strategy is more scalable, in the number of classes of the recognition problem, than those requiring simultaneous feature and classifier design.

The search for the optimal set of features, in the minimum BE sense, for a given classification problem can be addressed in two ways: by 1) *feature extraction* (FE) or 2) *feature selection* (FS). Denoting the *observation space* associated with the classification problem by $\mathcal{X}$ (typically high-dimensional), the goal of both FE and FS is to find the best transform $\mathbf{W}$ into a *feature space* $\mathcal{Y}$ (typically lower dimensional) where learning is easier. While in the case of FE there are few constraints on $\mathbf{W}$, for FS the transformation is constrained to be a projection, i.e. the components of a *feature vector* in $\mathcal{Y}$ are a subset of the components of the associated vector in $\mathcal{X}$. While both FS or FE can be used for the minimization of BE, both approaches have non-trivial limitations.

On one hand, FS requires the solution of a significantly simpler computational problem, since it consists of selecting the best subset from a set of already available basis functions. On the other, because it cannot produce features that are not part of the original set, the resulting transformation is usually sub-optimal. For example, two features that (as a pair) are highly discriminant but also highly correlated can have marginal distributions of small discriminant power. Such feature pairs cannot be reduced to a single new discriminant feature by FS techniques. FE avoids this problem by designing the basis itself, through the search for the overall optimal $\mathbf{W}$, but requires the solution of a significantly more difficult optimization problem. In fact, because BE is a non-linear function of the feature transformation, which does not have well-defined derivatives everywhere, its minimization by straightforward application of standard gradient-descent procedures can be quite challenging. Perhaps due to this, only a surprisingly small amount of work has addressed the direct minimization of BE in both the FE and FS literatures [12, 1].

In this paper, we introduce an algorithm for the computation of the minimum-BE feature set for a given classification problem. This algorithm combines the appealing properties of FS and FE. Like FS methods, it progresses in a sequence of steps where, at each step, the best features among those not yet selected are identified. However, unlike FS methods, it does not blindly include these features in the selected set. Instead, it considers the set of 2-D subspaces spanned by all pairs of features such that one feature is in the selected set and the other in the candidate set. It then performs FE in each of these subspaces, to find the direction that leads to the largest decrease in BE, and includes that direction in the selected set.

When compared to standard FE procedures, the new algorithm has the advantage of immediately zooming in on the optimal features that may already exist in the initial feature set. This leads to a significantly improved rate of convergence. When compared to FS procedures, it has the advantage of not being restricted to the original feature set. Experimental evaluation on multi-class visual recognition tasks shows that it converges to minimum Bayes error solutions in a very small number of iterations. The new algorithm is compared to the FE solutions in common use in the large-scale classification context - PCA, LDA and heteroscedastic discriminant analysis (HDA) [7] - and to an alternative FE solution based on gradient descent on a tight upper bound of the BE. It significantly outperforms these solutions, either by having much greater robustness to local minima or by achieving significantly faster convergence.

2. Minimum Bayes error features

Consider a set of training data $\{\mathbf{x}_l, c_l\}_{l=1}^N$ drawn from a continuous-valued random variable X such that $\mathbf{x}_l \in \Re^{n\times 1}$, and a discrete random variable C that generates class labels $c_l \in \{1, ..., |\mathcal{C}|\}$. The goal of FE is to find a linear feature transformation $\mathbf{W} : \mathcal{X} \subset \Re^{n\times 1} \to \mathcal{Y} \subset \Re^{m\times 1}$, $\mathbf{y}_l = \mathbf{W}\mathbf{x}_l$, $m < n$, that reduces the dimensionality of the data from n to m. The *minimum Bayes error* feature transformation $\tilde{\mathbf{W}}$ is the one that minimizes the *Bayes error* (BE)[6] on the output space $\mathcal{Y}$

$$L^*_{\mathcal{Y}} = 1 - E_Y\left[\max_c P_{C|Y}(c|\mathbf{y})\right] \quad (1)$$

$$= 1 - \int_{\Re^m} \max_c P_{C|Y}(c|\mathbf{y}) P_Y(\mathbf{y}) dy, \quad (2)$$

where $P_{C|Y}(c|\mathbf{y})$ is the posterior distribution for class c on $\mathcal{Y}$ and P_Y the probability density function for $\mathbf{y}$. Formally,

$$\tilde{\mathbf{W}} = \arg \min_{\mathbf{W}, rank(\mathbf{W})=m} L^*_{\mathcal{Y}}. \quad (3)$$

2.1. Estimating the Bayes error

Typically one does not have access to the probabilities $P_{C|Y}(c|\mathbf{y})$ or $P_Y(\mathbf{y})$ and it is therefore impossible to evaluate the BE through (2). Noting, however, that by the application of Bayes rule

$$L^*_{\mathcal{Y}} = 1 - E_Y\left[\max_c \frac{P_{Y|C}(\mathbf{y}_l|c)P_C(c)}{\sum_c P_{Y|C}(\mathbf{y}_l|c)P_C(c)}\right], \quad (4)$$

it follows that, given the class-conditional densities $P_{Y|C}(\mathbf{y}_l|c)$, the priors $P_C(c)$, and a sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, the expectation above can be estimated by the Monte-Carlo approximation

$$\hat{L}^*_{\mathcal{Y}} = 1 - \frac{1}{N}\sum_l \left[\max_c \frac{P_{Y|C}(\mathbf{y}_l|c)P_C(c)}{\sum_c P_{Y|C}(\mathbf{y}_l|c)P_C(c)}\right], \quad (5)$$

which we denote by the *empirical Bayes error* (EBE). The class priors are assumed known (but could also be estimated from training data quite easily), while the class-conditional

densities are estimated by maximum likelihood (via the expectation-maximization algorithm [5]) , using a Gaussian mixture model

$$P_{X|C}(\mathbf{x}_l|c) = \sum_{k=1}^{K_c} \lambda_{ck} \mathcal{G}(\mathbf{x}_l; \mu_{ck}, \mathbf{\Sigma}_{ck}) \quad (6)$$

in $\mathcal{X}$, and leading to a Gaussian mixture in $\mathcal{Y}$

$$P_{Y|C}(\mathbf{y}_l|c) = \sum_{k=1}^{K_c} \lambda_{ck} \mathcal{G}(\mathbf{W}\mathbf{x}_l; \mathbf{W}\mu_{ck}, \mathbf{W}\mathbf{\Sigma}_{ck}\mathbf{W}^T). \quad (7)$$

Note that this estimation is an initialization step that only has to be performed once, typically when the images in the class are added to the database, and is likely to be required for operations other than feature design (e.g. the actual classification of images presented to the recognition system). Hence, it does not affect the complexity of the feature design algorithms to be discussed in the subsequent sections.

2.2. Joint feature selection and extraction

The matrix $\mathbf{W}$ can be seen as the product of a matrix $\mathbf{W}_0$ whose rows form a basis of $\mathcal{X}$ and the canonical projection matrix $\Pi_n^m : \Re^n \rightarrow \Re^m$, $\Pi_n^m(x_1, \ldots, x_n) = (x_1, \ldots, x_m)$

$$\mathbf{W} = \Pi_n^m \mathbf{W}_0. \quad (8)$$

Under this interpretation, the rows of $\mathbf{W}$ are simply the subset of the basis vectors of $\mathcal{X}$ that span a subspace $\mathcal{X}_s \subset \mathcal{X}$. The BE on $\mathcal{Y}$ is determined by how discriminant this subspace is, i.e. it will be minimum when $\mathcal{X}_s$ is the most discriminant m-dimensional subspace of $\mathcal{X}$. Since discarding a discriminant direction can lead to a drastic increase in BE, the transformation $\mathbf{W}$ can be significantly improved by switching a basis vector of $\mathcal{X}_s^c$ (row-vectors of $\mathbf{W}_0$ not in $\mathbf{W}$) with a basis vector of $\mathcal{X}_s$ (i.e. row vectors of $\mathbf{W}$) when the former is a better discriminant than the latter.

This is the basic operation of FS, and one that is very unlikely under traditional FE. Because, when seen as points in $\Re^{n \times m}$, the matrices $\mathbf{W}$ before and after the switch can be arbitrarily far apart, it is highly likely that local minima of the BE surface will prevent a gradient-descent type of iteration from reaching the latter when initiated at the former. Due to this ability to avoid local optima (the step in solution space is not guided by the gradient) FS usually has a significantly faster convergence rate than FE. The only problem is that it can never identify discriminant directions which are not basis functions of $\mathbf{W}_0$ already. This can be a significant limitation, as illustrated by Figure 1. In this example, while the features x_1 and x_2 are (jointly) a highly discriminant pair, their marginal class-conditional densities exhibit a significant amount of overlap. Hence, because none of the two features is significantly discriminant by itself, it is unlikely that, in the context of a larger problem, the highly discriminant pair would be identified by a standard FS step.

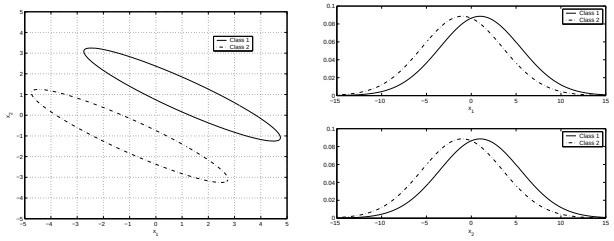


Figure 1. A classification problem with a pair of jointly discriminant features that, individually, are not very discriminant.

In order to achieve convergence rates equivalent to those of FS, while avoiding this limitation, we introduce an algorithm that performs joint FS and FE, which we denote by FSE (feature selection and extraction). The basic idea is to replace the simple evaluation of the goodness of the switch between the two candidate vectors with a full FE step in the plane spanned by them. Let $\mathbf{w}_i$ be the vector in $\mathcal{X}_s$ (the i^{th} row of $\mathbf{W}_0, i \in \{0, \ldots, m-1\}$) and $\mathbf{w}_o$ the one in $\mathcal{X}_s^c$ (o^{th} row of $\mathbf{W}_0, o \in \{m, \ldots, n-1\}$), and consider the set of 2D rotation matrices $R(i, o, \theta_{io})$ (where $R(i, o, \theta_{io})$ is identical to the $n \times n$ identity matrix with the exception of $R_{ii} = \cos(\theta_{io}), R_{io} = \sin(\theta_{io}), R_{oi} = -\sin(\theta_{io}), R_{oo} = \cos(\theta_{io})$). Instead of simply evaluating the EBE resulting from the switch of $\mathbf{w}_i$ with $\mathbf{w}_o$, we search for the rotation angle θ_{io} that leads to the overall transformation

$$\mathbf{W} = \Pi_n^m R(i, o, \theta_{io}) \mathbf{W}_0, \quad (9)$$

with smallest EBE

$$\hat{L}_{\mathcal{Y}}^* = 1 - \frac{1}{N} \sum_{l=1}^{N} \max_c P_{C|Y}(c|\mathbf{y}_l) \quad (10)$$

where $P_{C|Y}(c|\mathbf{y}_l)$ is obtained by combining (7) and the class priors with Bayes rule. This is a one dimensional minimization problem that can, therefore, be solved very efficiently with standard exhaustive search procedures (e.g. golden search [9]).

In fact, it is usually not even necessary to repeat this procedure for all possible pairs of basis vectors. One observation that we have made quite consistently is that, when $\mathbf{W}_0$ is a sensible initializion (e.g. that provided by PCA), the vast majority of the planes $(\mathbf{w}_i, \mathbf{w}_o)$ either 1) are not very discriminant, or 2) already have $\mathbf{w}_i$ as the most discriminant dimension. In these cases there is not much to be gained from the rotation and it is unlikely that such planes will be selected. To take advantage of this observation, we introduce an (optional) pre-filtering step that eliminates the planes with small ratio between 1) the EBE of the projection on $\mathbf{w}_i$

$$\tilde{L}_{[\mathbf{w}_i]}^* = 1 - E_X \left[\max_c P_{C|X}(c|\mathbf{w}_i \mathbf{x}) \right], \quad (11)$$

and 2) the EBE of the projection on the plane

$$\tilde{L}^*_{[\mathbf{w}_i,\mathbf{w}_o]} = 1 - E_X\left[\max_c P_{C|X}\left(c\middle|\begin{bmatrix}\mathbf{w}_i\\ \mathbf{w}_o\end{bmatrix}\mathbf{x}\right)\right]. \quad (12)$$

Note that, because all the densities involved are one or two-dimensional, this ratio can be computed using histogram-based density estimates.Its complexity is therefore negligible when compared to that of (10) and, if p planes are selected, the overall complexity is reduced by a factor of $sm(n-m)/p$. The complete algorithm is as follows:

1. let $\mathbf{W} = \Pi_n^m \mathbf{W}_0$;
2. compute $\frac{\tilde{L}^*_{[\mathbf{w}_i]}}{\tilde{L}^*_{[\mathbf{w}_i,\mathbf{w}_o]}}$ for all pairs $(\mathbf{w}_i, \mathbf{w}_o)$ and select the p pairs of smallest ratio.
3. for each of the p selected pairs find the rotation angle θ^*_{io}, using golden section search, that yields the smallest possible EBE as given by (9) and (10)
4. find the plane $(\mathbf{w}_{i^*}, \mathbf{w}_{o^*})$ that leads to the smallest empirical BE and update $\mathbf{W}_0 = R(i^*, o^*, \theta^*_{i^*o^*})\mathbf{W}_0$.
5. return to step 2 until the EBE difference between 2 successive iterations is smaller than a constant t (set to 10^{-6} in our experiments).

The matrix $\mathbf{W}_0$ can be the identity but can also be a feature transformation itself. One sensible solution is to rely on a feature transformation that experience has shown to perform reasonably well on the problem at hand. For example, a principal component analysis or a wavelet decomposition in visual recognition problems. In fact, as long as $\mathbf{W}_0$ is invertible, there will be no loss of BE and, therefore, any orthogonal or overcomplete decomposition qualifies.

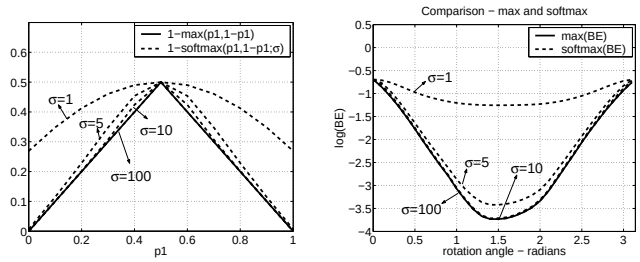
2.3. Gradient descent

As a benchmark against which to compare the algorithm of the previous section, we implemented an algorithm based on FE alone. As is customary in the FE literature, this algorithm performs gradient descent on the EBE surface. It turns out that the solution to this problem is not straightforward since, due to the $\max(\cdot)$ operator in (2), the EBE surface does not have well-defined derivatives everywhere. To overcome this limitation, we relied on the upper bound resulting from the replacement of the $\max(\cdot)$ operator by the *softmax* function

$$s(\{x_i\};\sigma) = \sum_j \frac{e^{\sigma x_j}}{\sum_i e^{\sigma x_i}} x_j, \quad (13)$$

where $\sigma > 0$ is a scale parameter, and $\{x_i\} \geq 0$ the input set [3]. As illustrated by Fig. 2 a), the softmax is a lower bound to the max function that can be made arbitrarily tight by taking σ to infinity. In practice, even for relatively small values of σ (e.g. $\sigma = 10$), the bound is a very good approximation to the max function. Consequently,

$$\hat{L}^*_{\mathcal{Y}} = 1 - E_Y\left[\sum_{c=1}^{|\mathcal{C}|} \frac{e^{\sigma P_{C|Y}(c|\mathbf{y})}}{\sum_{d=1}^{|\mathcal{C}|} e^{\sigma P_{C|Y}(d|\mathbf{y})}} P_{C|Y}(c|\mathbf{y})\right] \quad (14)$$



a) b)

Figure 2. The softmax function represents a tight bound of the $\max$ function.

is a very good approximation to (2). This is illustrated by Fig. 2 b), which presents the BE on a problem with $n = 2$, $m = 1$, $|\mathcal{C}| = 2$, as a function of the angle of the line into which the input space is projected (see Fig.3 a)). Clearly, the extrema of the two functions are co-located. Furthermore, because (14) has continuously differentiable derivatives, it can be minimized with standard gradient descent

$$\mathbf{W}_{(t+1)} = \mathbf{W}_{(t)} - \eta\left(\frac{\partial \hat{L}^*_{\mathcal{Y}}}{\partial \mathbf{W}}\right)_{(t)}, \quad (15)$$

where t represents the time step, and η is a learning rate (in our implementation the value that produces the largest decay of the cost among a set of pre-defined values). Replacing all expectations by the empirical means $E_Y[f(\mathbf{y})] = \frac{1}{N}\sum_l f(\mathbf{y}_l)$, it follows, after some algebraic manipulation[1], that

$$\frac{\partial \hat{L}^*_{\mathcal{Y}}}{\partial \mathbf{W}} \approx -\frac{1}{N}\sum_{l=1}^{N}\left[\sum_{c=1}^{|\mathcal{C}|} \frac{e^{\sigma P_{C|Y}(c|\mathbf{y}_l)}}{\sum_{d=1}^{|\mathcal{C}|} e^{\sigma P_{C|Y}(d|\mathbf{y}_l)}}\left(\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}}\right)\right] \quad (16)$$

where, by application of Bayes rule,

$$\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}} = \left[\frac{1}{P_Y(\mathbf{y}_l)}\left(\frac{\partial P_{Y|C}(\mathbf{y}_l|c)}{\partial \mathbf{W}}\right) P_C(c) - \left(\frac{P_{C|Y}(c|\mathbf{y}_l)}{P_Y(\mathbf{y}_l)}\right)\left(\frac{\partial P_Y(\mathbf{y}_l)}{\partial \mathbf{W}}\right)\right], \quad (17)$$

with

$$P_Y(\mathbf{y}_l) = \sum_{c=1}^{|\mathcal{C}|} P_{Y|C}(\mathbf{y}_l|c) P_C(c),$$

$$\frac{\partial P_Y(\mathbf{y}_l)}{\partial \mathbf{W}} = \sum_{c=1}^{|\mathcal{C}|} \frac{\partial P_{Y|C}(\mathbf{y}_l|c)}{\partial \mathbf{W}} P_C(c),$$

1 Assuming that $s\left(P_{C|Y}(c|\mathbf{y}_l)\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}};\sigma\right) \approx s\left(P_{C|Y}(c|\mathbf{y}_l)s\left(\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}};\sigma\right);\sigma\right)$, which is an equality when $s(\{.\},\sigma)$ is replaced by $\max(\{.\})$

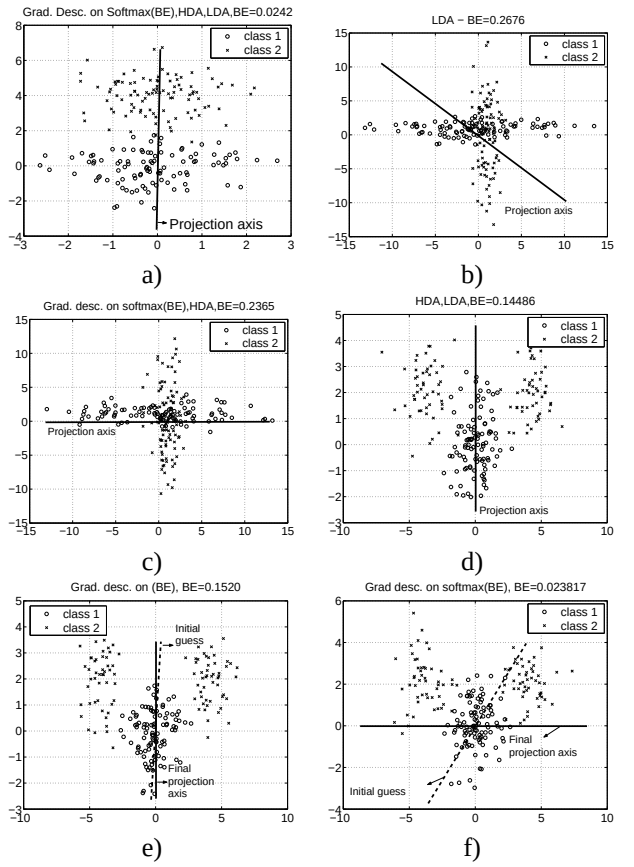


Figure 3. Various toy problems and the solutions obtained by LDA, HDA, and gradient descent. In all cases the best $1D$ subspace is represented by the solid bar.

and $P_C(c) = \frac{1}{|C|}$. Under the Gauss mixture assumption of (7)

$$\begin{aligned}\frac{\partial P_{Y|C}(\mathbf{y}_l|c)}{\partial \mathbf{W}} &= \frac{\partial P_{Y|C}(\mathbf{W}\mathbf{x}_l|c)}{\partial \mathbf{W}} \\ &= \sum_{k=1}^{K_c} \lambda_{ck}\Psi(c,k)\left(-\Omega(c,k) - \Gamma(c,k,\mathbf{x}_l)\right)\beta(c,k,\mathbf{x}_l),\end{aligned} \tag{18}$$

with

$$\begin{aligned}\Omega(c,k) &= (\mathbf{W}\Sigma_{ck}\mathbf{W}^T)^{-1}\mathbf{W}\Sigma_{ck} \\ \Psi(c,k) &= (2\pi)^{-\frac{m}{2}}\left|\mathbf{W}\Sigma_{ck}\mathbf{W}^T\right|^{-\frac{1}{2}} \\ \Gamma(c,k,\mathbf{x}_l) &= (\mathbf{W}\Sigma_{ck}\mathbf{W}^T)^{-1}\mathbf{W}(\mathbf{x}_l-\mu_{ck})(\mathbf{x}_l-\mu_{ck})^T \\ &\quad \left(I - \mathbf{W}^T(\mathbf{W}\Sigma_{ck}\mathbf{W}^T)^{-1}\mathbf{W}\Sigma_{ck}\right) \\ \beta(c,k,\mathbf{x}_l) &= e^{-\frac{1}{2}(\mathbf{W}(\mathbf{x}_l-\mu_{ck}))^T(\mathbf{W}\Sigma_{ck}\mathbf{W}^T)^{-1}(\mathbf{W}(\mathbf{x}_l-\mu_{ck}))}.\end{aligned}$$

Finally, the scale parameter is set to $\sigma = \arg\max_\sigma \left\| \frac{\partial \hat{L}^*_{\mathcal{Y}}}{\partial \mathbf{W}} \right\|$, i.e. the value that maximizes the gradient of the cost function.

3. Experiments

To evaluate the algorithms introduced in this work, we applied them to various visual recognition problems, ranging from simple toy examples that provide intuition to full-blown recognition tasks involving many classes in domains such as face or texture recognition. We started with a set of problems designed to compare the performance of the new algorithms to that of the classical solutions, namely LDA and HDA. The first set of experiments were performed on a collection of toy problems (projection of two classes from 2 to 1 dimension) that provide some intuition about the advantages of minimizing BE. As illustrated by Fig. 3 a), all methods performed perfectly on Gaussian problems with classes of equal covariance. However, as shown in b) and c), LDA broke down even for Gaussian problems of unequal class covariance. This is a well known problem and the motivation for HDA [7, 12]. Both HDA and the two minimum BE algorithms converged to the optimal solution, shown in c).

The problem on Figures 3 d)-3 f) consists of a Gaussian class and a second class which is a mixture of two Gaussians. In this case, the BE surface has a local minimum that, as shown in d), is also the optimal solution for LDA and HDA. Fig. 3 e) and f) illustrate the susceptibility of the gradient descent algorithm to local minima of the BE. As can be seen in e), if the initial $\mathbf{W}$ is close to a local minimum then gradient descent will converge to it. There is however, as shown in f), a much larger region of the solution space that will lead to convergence to the global minimum. Finally, this problem demonstrates the increased robustness of FSE to local minima. Because the optimal direction is found by exhaustive search we were not able to find, under FSE, an initialization that would prevent convergence to the global minimum.

The second set of experiments was performed on a face recognition task using the ORL database. This database contains 20 classes, each composed of 10 112×92 images, which were scaled down to 15×13 (by smoothing and bicubic interpolation). This set was split into a training database (first 8 images of each class) and a test database (remaining 2 images). The matrix $\mathbf{W}_0$ was the PCA matrix of the training data, as used in the popular eigenfaces technique [13], which was also used as the initial basis for HDA. Recognition was performed with a maximum likelihood classifier $g^*(\mathbf{W}\mathbf{x}_l) = \arg\max_c P_{Y|C}(\mathbf{W}\mathbf{x}_l|c)$, where $\mathbf{x}_l$ is a face from the test database, and $P_{Y|C}(\mathbf{y}|c)$ the Gaussian learned from the training images of class c. Table 1 shows the values of EBE obtained on the training and test sets when $\mathbf{W}$ is learned with PCA, HDA, and FSE algorithm. Besides the fact that FSE outperforms the two other techniques, it is interesting to notice the correlation between the Bayes error and the actual probability of error. This correlation is confirmed by Fig. 4, which presents a scatter plot of the two quantities, obtained by varying the transformation $\mathbf{W}$ and the output dimensionality m. This is particularly interesting given that the

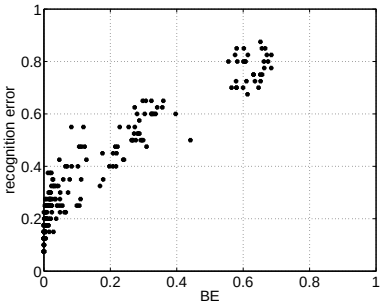


Figure 4. BE vs error rate on the ORL database.

classifier is sub-optimal (it is unlikely that the PCA features are exactly Gaussian) and there were, therefore, no guarantees that the recognition error would behave like the BE.

	FSE	PCA	HDA
BE training set	0.0011	0.0124	0.0015
BE testing set	0.0026	0.0575	0.19
Recognition rate	70%	66.7%	48.3 %

Table 1. BE and recognition rates on the ORL face recognition experiment.

Brodatz is also interesting in the sense that it poses a significant problem for many classification architectures. For example, the straightforward application of a support vector machine (SVM) to this database tends to perform quite poorly. Table 2 presents the best results that we were able to obtain, at several image resolutions, for an SVM with a Gaussian kernel, after a substantial amount of tuning of both the kernel variance and the SVM capacity parameter[2]. We

Resolution	Recognition rate
8×8	32.08
16×16	32.08
32×32	31.25
128×128	33

Table 2. Recognition rates on Brodatz for an SVM classifier at different image resolutions.

believe that this strongly disappointing performance is due to

2 We started from a kernel variance equal to the median Euclidean distance between the training vectors and a capacity of 1, and then manually tried various variations of the two parameters around these initial values. The combination that lead to smallest error was selected.

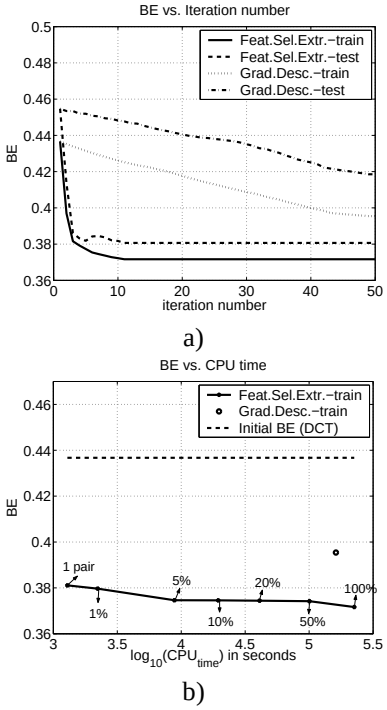


Figure 5. a) empirical BE vs number of iterations for gradient descent and FSE on the test and training datasets of Brodatz. b) empirical BE vs computational time required for convergence by FSE as a function of the parameter p (solid line), initial BE (dashed) and BE vs computational cost of gradient descent (dot).

the fact that the $1-vs-all$ strategy required to turn the multi-class problem (that the SVM cannot handle directly) into a collection of binary problems (which are then combined into a multi-class decision) may be strongly sub-optimal on Brodatz. We have also previously shown that other currently popular representations in learning and vision, e.g. an independent component analysis (ICA) type of decomposition, do not work well on this database [1]. In fact, an extensive study comparing the performance of various feature spaces (including PCA, ICA, and wavelets), have shown that the discrete cosine transform (DCT) is a top performer on Brodatz (see [1] for details). We therefore used the DCT as initial basis $\mathbf{W}_0$, in an attempt to determine if further optimization, by either FSE or gradient descent, could lead to visible improvement over this already very good solution.

We started by comparing the performance of the minimum-BE feature sets obtained by FSE and gradient descent, saving the matrix $\mathbf{W}$ at each iteration and measuring the corresponding EBE on both the training and test sets, to make sure that there was no over-fitting. Fig. 5 presents the evolution of the EBE as a function of the iteration number, showing that the convergence rate of FSE is significantly

faster (at least one order of magnitude) than that of gradient descent. By running the algorithms for an extended number of iterations, we also observed that the curves remained flat after 50 iterations. This means that gradient descent was trapped in a local minimum that prevented convergence to the better solution reached by FSE. In summary, gradient descent required a significantly larger number of iterations to converge to a worse solution than that found by FSE.

In order to compare the computational cost of the two algorithms (and evaluate the trade-off between BE and complexity due to the filtering step of FSE), we ran FSE with various values of the plane-retention parameter p. Fig. 5 b) shows the variation of the final value of BE, for $p = 1$ and $p \in \{1\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$ of all possible planes, as a function of the CPU time[3]. Also shown are the BE achieved by gradient descent and the corresponding time and the initial BE. Clearly, simply picking the best plane is enough to reach a solution that is very close to the best possible (and better than the gradient descent solution), at a computational cost more than two orders of magnitude smaller than that of either the overall best or gradient descent.

Finally, we compared the recognition performance of the FSE solution with that of the initial DCT features. Recognition was performed with a maximum likelihood classifier $g^*(\mathbf{W}\mathbf{x}_l) = \arg\max_c P_{Y|C}(\mathbf{W}\mathbf{x}_l|c)$, where $\mathbf{x}_l$ is an image from the test database, and $P_{Y|C}(\mathbf{y}|c)$ the Gaussian mixture learned from the training images of class c. Table 3 shows the recognition rates obtained, confirming that the FSE solution is the best one and reduces the error rate of the DCT features by about 12%. Given that the DCT features already perform very well for most test images, we believe that this improvement is significant.

Features	Recognition rate
DCT	92.92
FSE	93.75

Table 3. Recognition rates on Brodatz for a mixture classifier based on the DCT and FSE feature spaces.

In fact, visual inspection of the classification results obtained for each test image revealed no instances where FSE did worse than the DCT. On the contrary, FSE tends to improve performance for test images belonging to classes that are visually quite similar to other classes in the database. These are the most difficult images to classify and the results above suggest that, for 12% of them, FSE is helpful. Furthermore, we have noticed that this gain is not achieved at the cost of a loss of the generalization ability of the classifier. On the contrary, the FSE-based classifier appears to be more robust than the DCT-based counterpart and produces

3 Computer configuration: Intel Xeon processor at 2.4GHz with 4GB of memory.

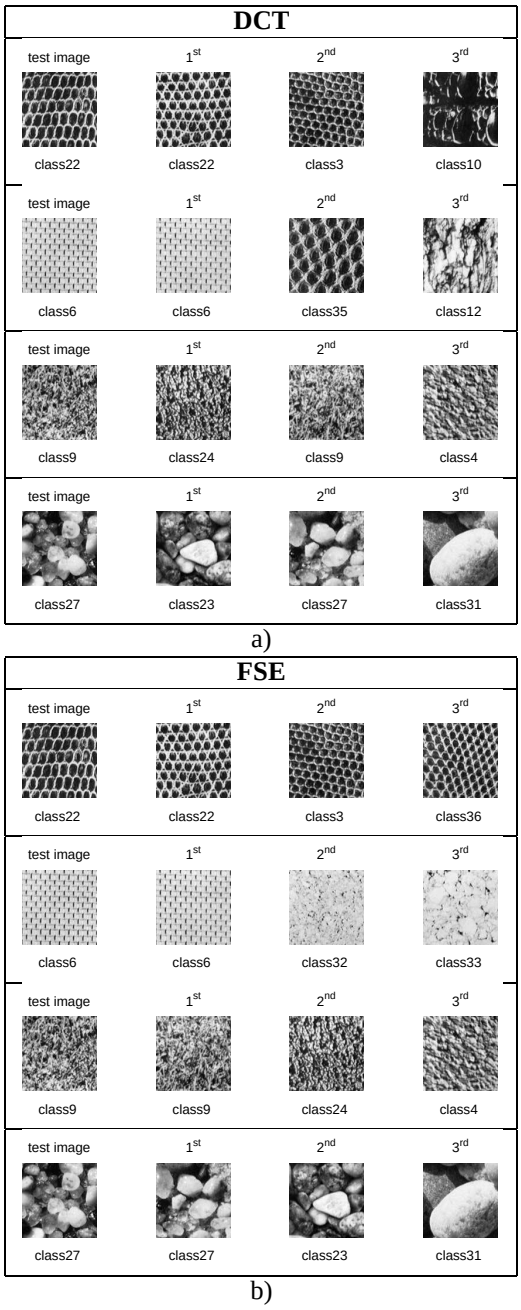


Figure 6. Recognition results obtained on Brodatz with the DCT-based (a) and FSE-based (b) classifiers. In each case, the classes in the database are ordered by decreasing likelihood with respect to the test image. For each class, we show a representative image.

judgments of similarity that seem more correlated to those of human perception. These points are illustrated by Figure 6, where we show the classification results obtained with the two classifiers for various test images. The top two examples of Figures 6 a) and 6 b) illustrate how the FSE-based classifier has better ability to generalize, producing an ordering of the classes that seems to be closer to human judgments of similarity. The bottom two examples of Figures 6 a) and 6 b) show instances where, even though close, the DCT-based classifier produces an error. In these cases, the FSE-based classifier is able to recover the correct ordering without altering the third match. All examples (as well as others that are omitted for brevity) support the argument that FSE produces a layout of the feature space that, locally, allows a finer discrimination between similar classes but, globally, brings those classes closer together.

Acknowledgements: We would like to thank Allan Jepson for useful suggestions during the preparation of this paper.

References

[1] N. Vasconcelos, G. Carneiro. What is the Role of Independence for Visual Recognition. *ECCV, Copenhagen, Denmark*, May 2002.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE PAMI*, 19(7):711–720, July 1997.

[3] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[4] Y. Le Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropogation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society*, B-39, 1977.

[6] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[7] N.Kumar and A.G.Andreou. Heteroscedastic discriminat analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26:283–297, 1998.

[8] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian Detection Using Wavelet Templates. In *IEEE CVPR, San Juan, Peurto Rico*, 1997.

[9] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, editors. *Numerical Recipes in C*. Cambridge University Press, 1992.

[10] D. Roth, M. Yang, and N. Ahuja. Learning to Recognize Three-Dimensional Objects. *Neural Computation*, 14:1071–1103, 2002.

[11] H. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE PAMI*, 20(1):23–38, January 1998.

[12] G. Saon and M. Padmanabhan. Minimum Bayes Error Feature Selection for Continuous Speech Recognition. In *NIPS*, Denver, USA, 2000.

[13] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 1991.

[14] P. Viola and M. Jones. Robust Real-Time Object Detection. In *Second International Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001.

[15] M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, pages 18–32, Dublin, Ireland, 2000.