Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 27 (2009) 131–140
www.elsevier.com/locate/imavis

Minimum Bayes error features for visual recognition

Gustavo Carneiro [a,*], Nuno Vasconcelos [b]

[a] Siemens Corporate Research, Integrated Data Systems Department, 755 College Road East, Princeton, NJ 08540, USA
[b] Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, Mail code 0407, EBU 1, Room 5603, La Jolla, CA 92093-0407, USA

Abstract

The design of optimal feature sets for visual classification problems is still one of the most challenging topics in the area of computer vision. In this work, we propose a new algorithm that computes optimal features, in the minimum Bayes error sense, for visual recognition tasks. The algorithm now proposed combines the fast convergence rate of feature selection (FS) procedures with the ability of feature extraction (FE) methods to uncover optimal features that are not part of the original basis function set. This leads to solutions that are better than those achievable by either FE or FS alone, in a small number of iterations, making the algorithm scalable in the number of classes of the recognition problem. This property is currently only available for feature extraction methods that are either sub-optimal or optimal under restrictive assumptions that do not hold for generic imagery. Experimental results show significant improvements over these methods, either through much greater robustness to local minima or by achieving significantly faster convergence.

Keywords: Visual recognition; Feature selection; Feature extraction; Minimum Bayes error; Mixture models; Face recognition; Texture recognition; Object recognition

1. Introduction

The fundamental goal for the design of a statistical classifier is to minimize its probability of making mistakes. More formally, if the classifier operates on observations from a random variable $\mathbf{X}$, defined on a *observation space* $\mathcal{X}$, and C is the random variable from which class labels are drawn, the goal is to design the decision function $g(\mathbf{x})$ that minimizes the probability of classification error [1].

$$E_{\mathbf{X}}[P_{C|\mathbf{X}}(c \neq g(\mathbf{x})|\mathbf{x})] = E_{\mathbf{X}}[1 - P_{C|\mathbf{X}}(g(\mathbf{x})|\mathbf{x})] \quad (1)$$

with $P_{C|X}(c|x)$ denoting the posterior distribution of class c, and $P_X(x)$ the marginal distribution of $\mathbf{X}$. The implementation of the minimum probability of error classifier requires access to (1) ideal estimates of the class conditional distributions $P_{X|C}(x|c)$, and (2) the space that best separates the two classes. In practice, the accuracy of the estimates $P_{X|C}(x|c)$ is affected by several factors (e.g., model assumed for the distributions, the accuracy with which model parameters can be learned from the available training data, etc.), but better estimates are usually obtained in low-dimensional spaces (where enough training data is more likely to be available). On the other hand, it can be shown that the space where classification takes place uniquely determines the lowest probability of error achievable by any classifier. Unlike the estimation error, this lower bound, known as the *Bayes error* (BE), tends to decrease with the dimension of the space [2]. Hence, the design of the optimal space for classification usually requires the identification, among all spaces that are low-dimensional enough to guarantee small density estimation error, that which achieves the minimum BE.

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

* Corresponding author. Tel.: +1 609 734 3570.
E-mail addresses: gustavo.carneiro@siemens.com (G. Carneiro), nuno@ece.ucsd.edu (N. Vasconcelos).
URLs: http://www.cs.ubc.ca/~carneiro (G. Carneiro), http://www.svcl.ucsd.edu/~nuno (N. Vasconcelos).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.06.008

(2) *feature selection* (FS). In both cases, the goal is to find the best transform **W** into a lower dimensional *feature space* $\mathcal{Y}$. While in the case of FE there are few constraints on **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, as illustrated in Fig. 1, 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 **W**, but requires the solution of a significantly more difficult optimization problem. In fact, because the 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 optimization 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 [3–5].

The most successful visual classification approaches in the literature find the optimal feature set for a given classification task by explicitly optimizing the performance of the classifier, thus skipping the estimation of the class conditional distribution $P_{X|C}(x|c)$. 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 [6,7], the selection of a best set of basis functions from a predefined set [8,9], or the selection of feature configurations [10,11], the end product is invariably a set of features that is optimal, in the classification sense, for recognition. The most challenging issue faced by such approaches is their 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. back-propagation, 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) [12], or linear discriminant analysis (LDA) [13], remain the most popular for problems such as face, object, or texture recognition.

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

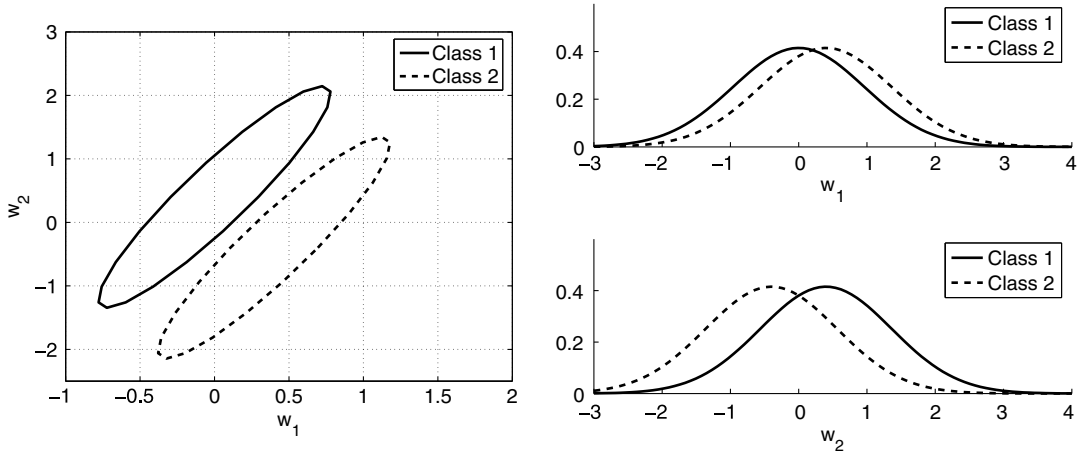

Fig. 1. A classification problem with a pair of jointly discriminant features that, individually, are not very discriminant.

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) [14] – 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 $\mathbf{x}_l \in \mathbb{R}^{n\times 1}$, and a discrete random variable C that generates class labels $c_l \in \{1,\ldots,|\mathscr{C}|\}$. The goal of FE is to find a feature transformation $f : \mathscr{X} \subset \mathbb{R}^{n\times 1} \to \mathscr{Y} \subset \mathbb{R}^{m\times 1}$. In this work, we consider $f(x)$ to be a linear function of x, i.e., $\mathbf{y}_l = f(\mathbf{x}_l)$ is written as $\mathbf{y}_l = \mathbf{W}\mathbf{x}_l$, that reduces the dimensionality of the data from n to m (i.e., $m < n$), The minimum BE feature transformation $\tilde{\mathbf{W}}$ is the one that minimizes the BE [1] on the output space $\mathscr{Y}$

$$\begin{aligned} L^*_{\mathscr{Y}} &= 1 - \int_{\mathbb{R}^m} \max_c P_{C|Y}(c|\mathbf{y}) P_Y(\mathbf{y})\,\mathrm{d}y \\ &= 1 - E_Y[\max_c P_{C|Y}(c|\mathbf{y})], \end{aligned} \tag{2}$$

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

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

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

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

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

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

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 [15]), using a Gaussian mixture model

$$P_{X|C}(\mathbf{x}|c) = \sum_{k=1}^{K_c} \lambda_{ck}\mathscr{G}(\mathbf{x};\boldsymbol{\mu}_{ck},\boldsymbol{\Sigma}_{ck}), \tag{6}$$

in $\mathscr{X}$, where

$$\mathscr{G}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Using well-known properties of the Gaussian, it can be easily shown that this leads to a Gaussian mixture in $\mathscr{Y}$ [2]

$$P_{Y|C}(\mathbf{y}|c) = \sum_{k=1}^{K_c} \lambda_{ck}\mathscr{G}(\mathbf{W}\mathbf{x};\mathbf{W}\boldsymbol{\mu}_{ck},\mathbf{W}\boldsymbol{\Sigma}_{ck}\mathbf{W}^{\mathrm{T}}). \tag{7}$$

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 $\mathscr{X}$ and the canonical projection matrix $\prod_n^m : \mathbb{R}^n \to \mathbb{R}^m$, $\prod_n^m(x_1,\ldots,x_n) = (x_1,\ldots,x_m)$

$$\mathbf{W} = \prod_n^m \mathbf{W}_0. \tag{8}$$

Under this interpretation, the rows of $\mathbf{W}$ are simply the subset of the basis vectors of $\mathscr{X}$ that span a subspace $\mathscr{X}_s \subset \mathscr{X}$. The BE on $\mathscr{Y}$ is determined by how discriminant this subspace is, i.e., it will be minimum when $\mathscr{X}_s$ is the most discriminant m-dimensional subspace of $\mathscr{X}$. 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 $\mathscr{X}_s^c$ (row-vectors of $\mathbf{W}_0$ not in $\mathbf{W}$) with a basis vector of $\mathscr{X}_s$ (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 $\mathbb{R}^{n\times m}$, 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 optimum (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 Fig. 1. In this example, while the features w_1 and w_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.

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, and 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 $\mathcal{X}_s^c$ (oth 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

$$\mathbf{W} = \prod_{n}^{m} R(i, o, \theta_{io}) \mathbf{W}_0 \tag{9}$$

with smallest EBE

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

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 [16]).

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 initialization (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 $\mathbf{w}_i$

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

and (2) the EBE of the projection on the plane

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

Note that, because all the densities involved are one- or two-dimensional, this ratio can be computed using histograms[1]. 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 $\mathbf{W} = \prod_{n}^{m} \mathbf{W}_0$;
(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 (7), (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. Fig. 2 illustrates the FSE steps on a problem where the goal is to reduce the dimensionality from 3 to 2, i.e., to find the plane, in 3D space, that best separates the two classes. In this case, the initial transformation is the identity and $\mathcal{X}_s$ is initially the plane spanned by $\mathbf{w}_1$ and $\mathbf{w}_2$. Note that there is substantial overlap between the projection of the class densities on this plane. FSE searches for a more discriminant plane as follows:

(1) Search for pairs of features $(\mathbf{w}_i, \mathbf{w}_o)$ such that $\mathbf{w}_i \in \mathcal{X}_s$, and $\mathbf{w}_o \in \mathcal{X}_s^c$, that span a subspace where the classes are better separated. In the example, since $\mathbf{w}_3$ is the only vector in $\mathcal{X}_s^c$, the possible combinations are $(\mathbf{w}_1, \mathbf{w}_3)$ and $(\mathbf{w}_2, \mathbf{w}_3)$.

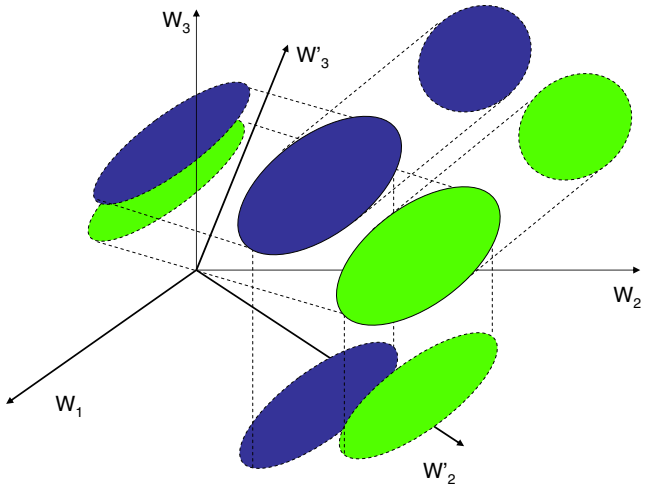

Fig. 2. Illustrative example showing the feature selection/extraction.

[1] In all experiments, we used histograms of bin size set according to the standard deviation of the projected data (for 1D, $\text{bin}_{\text{size}} = 6\frac{\text{std}(\mathbf{w}_i \mathbf{x}_l | l \in \{1,\ldots,N\})}{\#\text{bins}}$, where #bins is fixed).

(2) Find 2D projection that maximizes the ratio $\frac{\tilde{L}^*_{[w_i]}}{\tilde{L}^*_{[w_i,w_o]}}$. In the example, this is the space spanned by $(\mathbf{w}_2, \mathbf{w}_3)$, where the classes are best separated.
(3) Rotate in the plane $(\mathbf{w}_2, \mathbf{w}_3)$ by the rotation angle that minimizes the EBE in the output space: $\tilde{L}^*_{\mathbf{W}} = 1 - \frac{1}{N}\sum_{l=1}^{N} \max_d P_{C|Y}(c|\mathbf{y}_l)$. This produces the basis $(\mathbf{w}'_2, \mathbf{w}'_3)$. Projecting the classes on the plane spanned by this basis leads to a classification problem of very small EBE, since the projections are well separated.

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(·) 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(·) operator by the *softmax* function

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

where $\sigma > 0$ is a scale parameter, and $\{x_i\} \geqslant 0$ the input set [17]. As illustrated by Fig. 3a, the bound can be made arbitrarily tight by taking σ to infinity, but is a very good approximation to the max function even for relatively small values of σ (e.g., $\sigma = 10$). Consequently,

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

is a very good approximation to (2). This is illustrated by Fig. 3b, which presents the EBE on a problem with $n = 2$, $m = 1$, $|\mathscr{C}| = 2$, as a function of the angle of the line into which the input space is projected (see Fig. 4a). 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

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

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). The partial derivative of (14) with respect to $\mathbf{W}$ is then written as

$$\frac{\partial \hat{L}^*_{\mathscr{Y}}}{\partial \mathbf{W}} = -E_Y\left[\sum_{c=1}^{|\mathscr{C}|} \frac{\partial}{\partial \mathbf{W}}\left(\frac{e^{\sigma P_{C|Y}(c|\mathbf{y})}}{\sum_{d=1}^{|\mathscr{C}|} e^{\sigma P_{C|Y}(d|\mathbf{y})}} P_{C|Y}(c|\mathbf{y})\right)\right]. \tag{16}$$

Eq. (16) can be rewritten as follows:

$$\begin{aligned}\frac{\partial \hat{L}^*_{\mathscr{Y}}}{\partial \mathbf{W}} = -E_Y\Bigg[&s\left(\left\{\sigma P_{C|Y}(c|\mathbf{y})\frac{\partial P_{C|Y}(c|\mathbf{y})}{\partial \mathbf{W}}\right\};\sigma\right)\\ &-s\left(\left\{\sigma P_{C|Y}(c|\mathbf{y}) s\left(\left\{\frac{\partial P_{C|Y}(c|\mathbf{y})}{\partial \mathbf{W}}\right\};\sigma\right)\right\};\sigma\right) + s\left(\left\{\frac{\partial P_{C|Y}(c|\mathbf{y})}{\partial \mathbf{W}}\right\};\sigma\right)\Bigg],\end{aligned} \tag{17}$$

where $s(\{\cdot\};\sigma)$ is the softmax function (13). It should be clear from (16) that $s(\{\cdot\};\sigma)$ is the softmax function of the values in the set $\{\cdot\}$ across the classes $c \in \{1,\ldots,|\mathscr{C}|\}$. Replacing the expectation in (17) by the empirical mean $E_Y[f(\mathbf{y})] = \frac{1}{N}\sum_l f(\mathbf{y}_l)$, and assuming that

$$\begin{aligned}&s\left(\left\{\sigma P_{C|Y}(c|\mathbf{y}_l)\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}}\right\};\sigma\right)\\ &\quad\approx s\left(\left\{\sigma P_{C|Y}(c|\mathbf{y}_l) s\left(\left\{\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}}\right\};\sigma\right)\right\};\sigma\right),\end{aligned}$$

which is an equality when $s(\{\cdot\};\sigma)$ is replaced by $\max(\{\cdot\})$, it follows that

$$\frac{\partial \hat{L}^*_{\mathscr{Y}}}{\partial \mathbf{W}} \approx -\frac{1}{N}\sum_{l=1}^{N}\left[\sum_{c=1}^{|\mathscr{C}|} \frac{e^{\sigma P_{C|Y}(c|\mathbf{y}_l)}}{\sum_{d=1}^{|\mathscr{C}|} e^{\sigma P_{C|Y}(d|\mathbf{y}_l)}}\left(\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}}\right)\right], \tag{18}$$

where, by application of Bayes rule,

$$\frac{\partial P_{C|Y}(c|\mathbf{y}_l)}{\partial \mathbf{W}} = \left[\frac{1}{P_Y(\mathbf{y}_l)}\left(\frac{\partial P_{Y|C}(\mathbf{y}_l|c)}{\partial \mathbf{W}}\right)P_C(c) - \left(\frac{P_{C|Y}(c|\mathbf{y}_l)}{P_Y(\mathbf{y}_l)}\right)\left(\frac{\partial P_Y(\mathbf{y}_l)}{\partial \mathbf{W}}\right)\right], \tag{19}$$

with

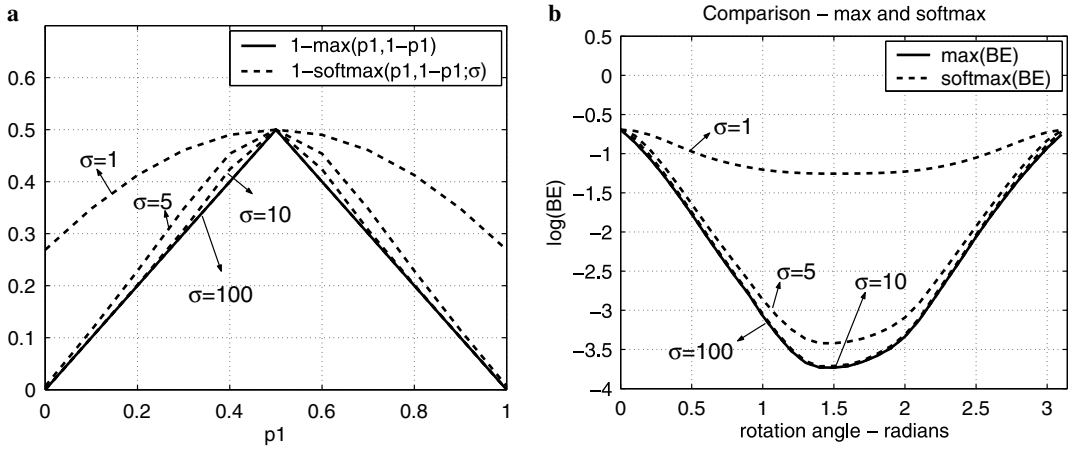

Fig. 3. The softmax function is a tight bound of the max function.

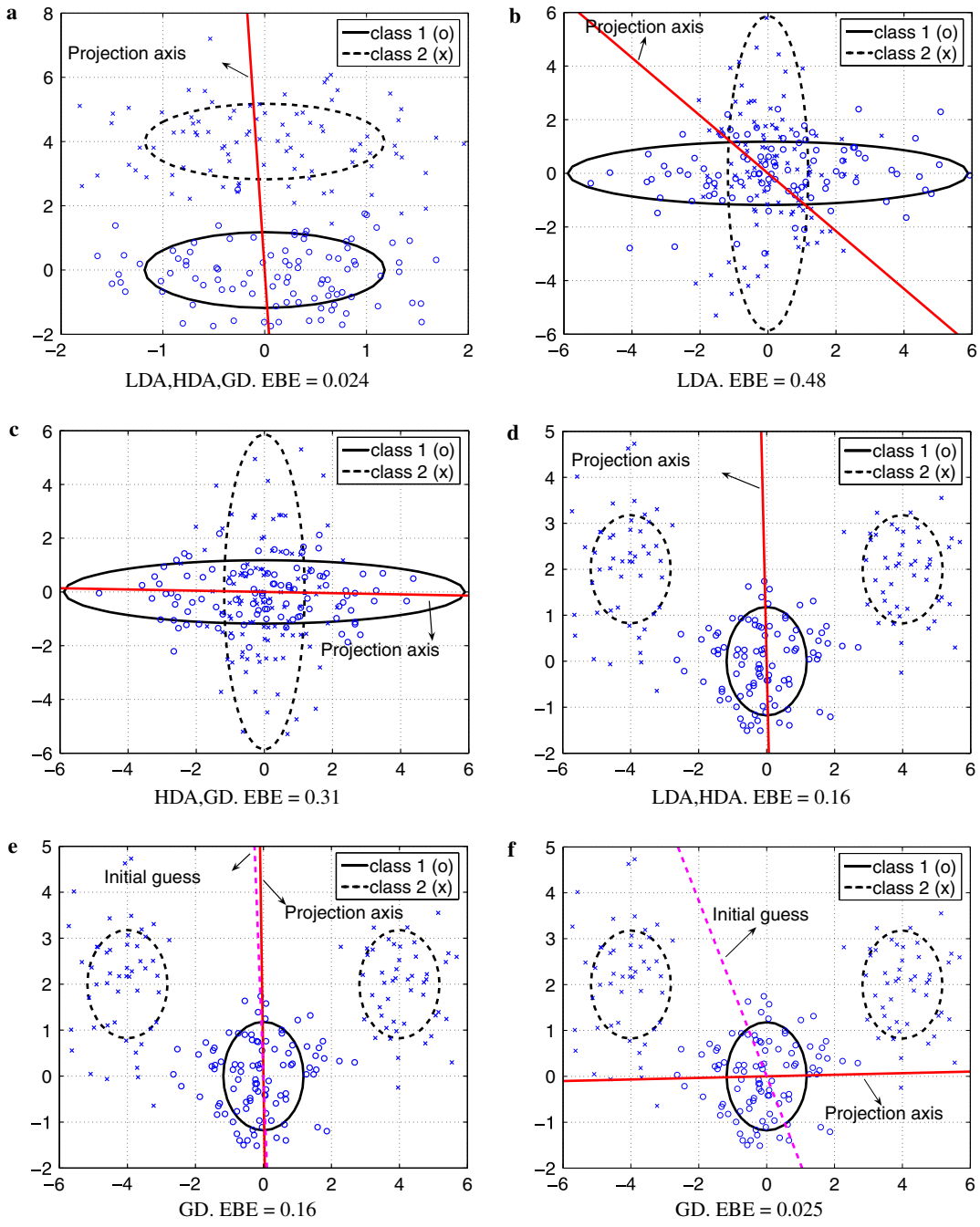

Fig. 4. Various toy problems and the solutions obtained by LDA, HDA, and gradient descent (GD). In all cases the best $1D$ subspace is represented by the solid bar, and the final value of the estimated Bayes error is shown at the bottom of the plot. Also, the methods used to obtain the solution depicted in each plot are listed at the bottom of the figure.

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

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

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

$$\frac{\partial P_{Y|C}(\mathbf{y}_l|c)}{\partial \mathbf{W}} = \frac{\partial P_{Y|C}(\mathbf{W}\mathbf{x}_l|c)}{\partial \mathbf{W}} = \sum_{k=1}^{K_c} \lambda_{ck} \Psi(c,k)(-\Omega(c,k) - \Gamma(c,k,\mathbf{x}_l))\beta(c,k,\mathbf{x}_l), \quad (20)$$

with

$$\Omega(c,k) = (\mathbf{W}\Sigma_{ck}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\Sigma_{ck},$$

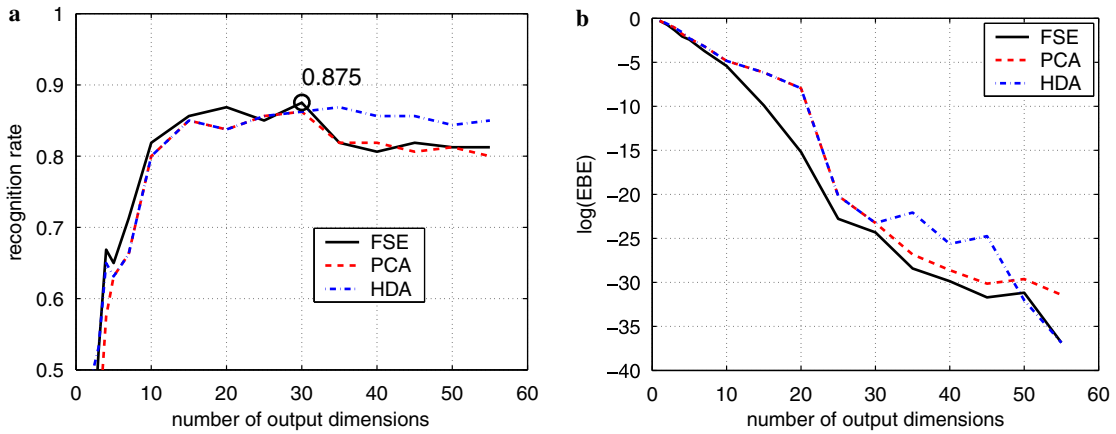

Fig. 5. Recognition results in the ORL database. The first graph (a) shows the recognition performance as a function of the number of dimensions of the output space computed using FSE, PCA and HDA. Graph (b) illustrates the log(EBE) for each space.

$$\Psi(c,k) = (2\pi)^{-\frac{m}{2}}|\mathbf{W}\Sigma_{ck}\mathbf{W}^{\mathrm{T}}|^{-\frac{1}{2}},$$

$$\Gamma(c,k,\mathbf{x}_l) = (\mathbf{W}\Sigma_{ck}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}(\mathbf{x}_l-\mu_{ck})(\mathbf{x}_l-\mu_{ck})^{\mathrm{T}} \times (I-\mathbf{W}^{\mathrm{T}}(\mathbf{W}\Sigma_{ck}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\Sigma_{ck}),$$

$$\beta(c,k,\mathbf{x}_l) = \mathrm{e}^{-\frac{1}{2}(\mathbf{W}(\mathbf{x}_l-\mu_{ck}))^{\mathrm{T}}(\mathbf{W}\Sigma_{ck}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}(\mathbf{x}_l-\mu_{ck}))}.$$

Finally, the scale parameter is set to $\sigma = \arg\max_\sigma \| \frac{\partial \hat{L}^*_{\sigma}}{\partial \mathbf{W}} \|$, 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 three classification problems, in which their performance was compared to that of the classical solutions. 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 EBE. Because in a 2D space the FSE algorithm performs an exhaustive search over all possible subspace (line) directions, we were not able to find any example, or initialization, that would prevent convergence to the global minimum. This was, however, not the case for most of the other techniques that we considered, namely LDA, HDA, or even the minimization of the EBE by gradient descent.

As illustrated by Fig. 4a, all methods performed well on Gaussian problems with classes of equal covariance. However, as shown in Fig. 4b and c, LDA broke down even for Gaussian problems of unequal class covariance. This is a well-known problem and the motivation for HDA [14,3]. Both HDA and the two minimum BE algorithms converged to the optimal solution, shown in Fig. 4c. The problem on Fig. 4d–f consists of a Gaussian class and a second class which is a mixture of two Gaussians. In this case, the EBE surface has a local minimum that, as shown in Fig. 4d, is also the optimal solution for LDA and HDA. Fig. 4e and f illustrate the susceptibility of the gradient descent algorithm to local minima of the EBE. As can be seen in Fig. 4e, if the initial $\mathbf{W}$ is close to a local minimum then gradient descent will converge to it. There is however, as shown in Fig. 4f, a much larger region of the solution space that will lead to convergence to the global minimum. This example is more illustrative of the problems faced by the minimization of EBE on high-dimensional spaces, where there can be many local minima. It demonstrates the increased robustness of FSE to this problem.

The second set of experiments was performed on a face recognition task using the ORL database. This database contains 40 classes, each composed of ten 112×92 images, which were scaled down to 15×13 (by smoothing and bicubic interpolation). This set was split into a training database (first 6 images of each class) and a test database (remaining 4 images). The matrix $\mathbf{W}_0$ was the PCA matrix of the training data, as used in the popular eigenfaces technique [12], 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. Note that the classes are assumed to be Gaussian, an assumption that favors HDA.

Fig. 5a shows the recognition rates, as a function of the number of output dimensions, obtained with FSE, PCA, and HDA. Note that the feature transform with 30 output dimensions computed by FSE holds the best overall recognition result of 87.50% (the best result recognition result for PCA was 86.25%, and for HDA 86.88%). Fig. 5b shows the EBE in the output space as a function of the output dimension, for each algorithm. Two conclusions can be drawn from this graph: (a) FSE produces the output space with minimum EBE for all dimensions, and (b) for all transforms, the EBE decreases with increasing dimensionality. Finally, Fig. 6 depicts the positive correlation between the EBE and the recognition error, on the experiments of Fig. 5a and b. This correlation is important, in the sense that it validates the claim that the minimization of EBE is a suitable criteria for optimal feature extraction. In particular, it shows that the FSE algorithm is in fact

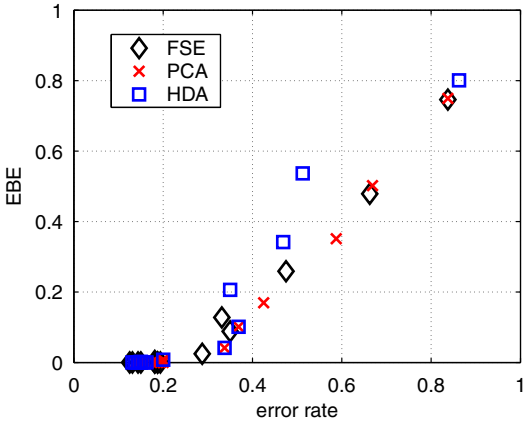

Fig. 6. Positive correlation between EBE and error rate.

Table 1
Recognition rates on Brodatz for an SVM classifier at different image resolutions

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

minimizing the classification error. Note that, unlike neural networks, SVMs, or boosted perceptrons this is achieved without the need to design the classifier at each iteration of the feature extraction process.

The next experiment is on the Brodatz texture database. Brodatz is 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 1 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 believe that this disappointing performance is due to 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 [18]. In fact, an extensive study comparing the performance of various feature spaces (including PCA, ICA, and wavelets), has shown that the discrete cosine transform (DCT) is a top performer on Brodatz (see [18] 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-EBE 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. 7 presents the evolution of the EBE as a function of the iteration number, showing that the convergence 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. 7b shows the variation of the final value of EBE, 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 EBE achieved by gradient descent and the corresponding time and the initial EBE. 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 2 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.

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

[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.

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

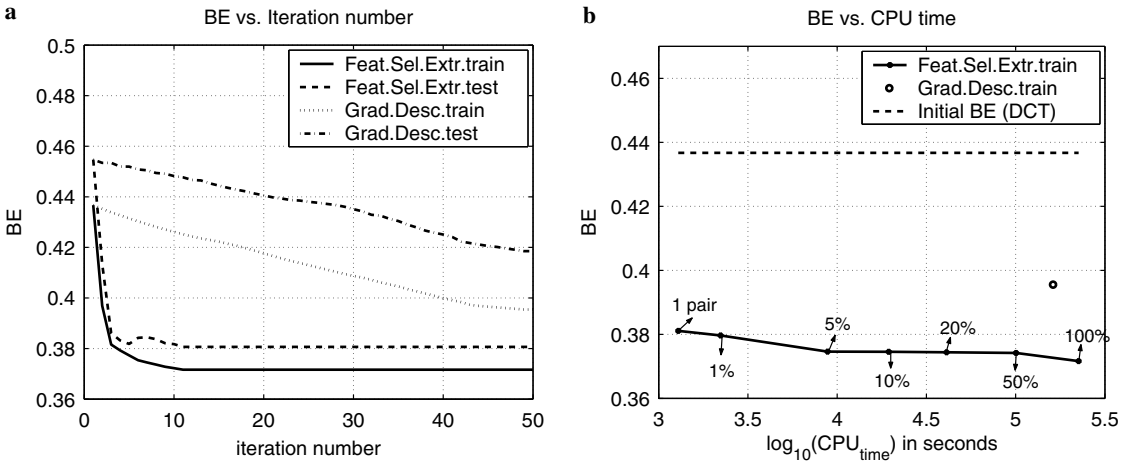

Fig. 7. (a) Empirical BE vs number of iterations for gradient descent and FSE on the test and training datasets of Brodatz. (b) Empirical BE vs computational time required for convergence by FSE as a function of the parameter p (solid line), initial EBE (dashed) and EBE vs computational cost of gradient descent (dot).

Table 2
Recognition rates on Brodatz for a mixture classifier based on the DCT and FSE feature spaces

Features	Recognition rate
DCT	92.92
FSE	93.75

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 judgments of similarity that seem more correlated to those of human perception. These points are illustrated by Fig. 8, where we show the classifi-

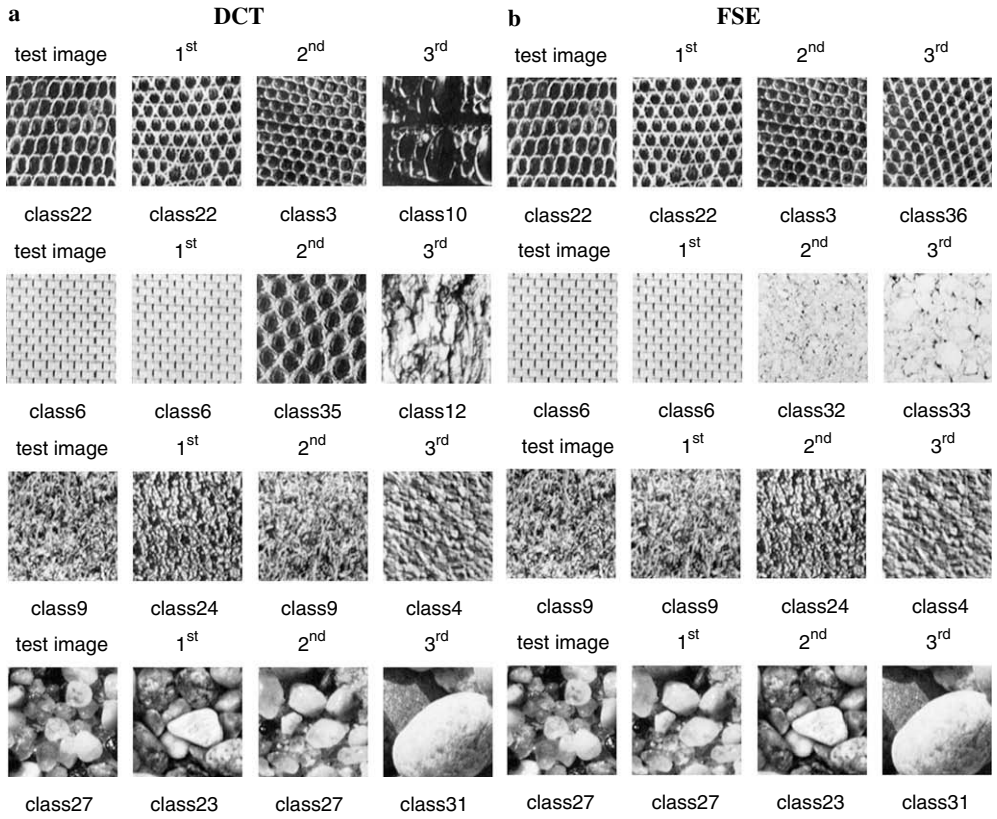

Fig. 8. Recognition results obtained on Brodatz with the DCT-based (a) and FSE-based (b) classifiers. In each case, the classes in the database are ordered by decreasing likelihood with respect to the test image. For each class, we show a representative image.

cation results obtained with the two classifiers for various test images. The top two examples of Fig. 8a and 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 Figs. 8a and 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.

4. Conclusion

We presented an algorithm to efficiently compute an optimal set of features in the minimum Bayes error sense. The algorithm combines the efficiency of feature selection with the ability of feature extraction to compute optimal features that are not part of the original set of basis functions. One important aspect of this work is that the feature set built by the algorithm now proposed can be used to train any type of classifier, thus separating the complex tasks of feature and classifier design. The divorce of these two tasks is important for large-scale classification problems not only in terms of efficiency, but also with respect to recognition accuracy. We provided empirical examples of the efficiency and efficacy of the algorithm in several visual recognition problems, ranging from simple toy examples to full-blown recognition tasks involving many classes in the domains of face and texture recognition.

Acknowledgements

We would like to thank Allan Jepson for useful suggestions during the preparation of this paper. This work was partially supported by NSF Career Grant IIS-0448609. Gustavo Carneiro was also partially funded by NSERC (Canada).

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[2] N. Vasconcelos, Minimum Probability of Error Image Retrieval, IEEE Trans. on Signal Processing 52 (8) (2004).
[3] G. Saon, M. Padmanabhan, Minimum Bayes Error Feature Selection for Continuous Speech Recognition, in: Proc. Neural Information Proc. Systems, Denver, USA, 2000.
[4] N. Vasconcelos, Feature selection by maximum marginal diversity, in: Neural Information Processing Systems, Vancouver, Canada, 2002.
[5] N. Vasconcelos, Feature Selection by Maximum Marginal Diversity: Optimality and Implications for Visual Recognition, in: Proc. IEEE Computer Vision and Pattern Recognition Conf., Madison, Wisconsin, 2003.
[6] Y.L. Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Backpropogation applied to handwritten zip code recognition, Neural Computation 1 (1989) 541–551.
[7] H. Rowley, S. Baluja, T. Kanade, Neural Network-Based Face Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.
[8] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian Detection Using Wavelet Templates, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Peurto Rico, 1997.
[9] P. Viola, M. Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, in: IEEE Conf. on Computer Vision and Pattern Recognition, Kawai, Hawaii, 2001.
[10] D. Roth, M. Yang, N. Ahuja, Learning to recognize three-dimensional objects, Neural Computation 14 (2002) 1071–1103.
[11] M. Weber, M. Welling, P. Perona, Unsupervised Learning of Models for Recognition, in: European Conference on Computer Vision, Dublin, Ireland, 2000, pp. 18–32.
[12] M. Turk, A. Pentland, Eigenfaces for Recognition, Journal of Cognitive Neuroscience 3 (1991).
[13] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[14] N. Kumar, A.G. Andreou, Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition, Speech Communication 26 (1998) 283–297.
[15] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society B-39 (1977).
[16] W. Press, S. Teukolsky, W. Vetterling, B. Flannery (Eds.), Numerical Recipes in C, Cambridge University Press, 1992.
[17] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, 1995.
[18] N. Vasconcelos, G. Carneiro, What is the Role of Independence for Visual Regognition?, in: Proc. European Conference on Computer Vision, Copenhagen, Denmark, 2002.