Appears in IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin, 2003.

Feature Selection by Maximum Marginal Diversity: optimality and implications for visual recognition

Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
nuno@ece.ucsd.edu

Abstract

We have recently shown that 1) the infomax principle for the organization of perceptual systems leads to visual recognition architectures that are nearly optimal in the minimum Bayes error sense, and 2) a quantity which plays an important role in infomax solutions is the marginal diversity(MD): the average distance between the class-conditional density of each feature and their mean. Since MD is a discriminant quantity and can be computed with great efficiency, the principle of maximum marginal diversity (MMD) was suggested for discriminant feature selection. In this paper, we study the optimality (in the infomax sense) of the MMD principle and analyze its effectiveness for feature selection in the context of visual recognition. In particular, 1) we derive a close form relation between the optimal infomax and MMD solutions, and 2) show that there is a family of classification problems for which the two are identical. Examination of this family in light of recent studies on the statistics of natural images suggests that the equivalence conditions are likely to hold for the problem of visual recognition. We present experimental evidence supporting the conclusions that 1) MD is a good predictor for the recognition ability of a given set of features, 2) MMD produces features that are more discriminant than those obtained with currently predominant criteria such as energy compaction, and 3) the extracted features are detectors of visual attributes that are perceptually relevant for low-level image classification.

1 Introduction

It has long been recognized that a good selection of visual measurements, usually known as *features*, is an important requirement for successful recognition systems. Given a feature space $\mathcal{Z}$, the goal of feature selection is to find the best projection T into a lower dimensional subspace $\mathcal{X}$ where learning is easier (e.g. can be performed with less training data). The only constraint on T is that the components of a *feature vector* in $\mathcal{X}$ are a subset of the components of the associated vector in $\mathcal{Z}$. Formally, this problem can be formulated as an optimization task, where the objective is to find the projection matrix that best satisfies a given criteria for "feature goodness".

In the context of visual recognition, various such criteria have been proposed throughout the years, the most popular of which is arguably *energy compaction* [12, 8], i.e. the best features are those that contain the largest fraction of the total energy. However, adopting the energy compaction principle neglects the fact that, for recognition, the best feature spaces are those that maximize *discrimination*, i.e. the separation between the different image classes to recognize. This has motivated vision researchers to revisit classical discriminant criteria, such as the ratio of between to within-class scatter behind classical *linear discriminant analysis* [2]. While an improvement over energy compaction, such criteria make very specific assumptions regarding class densities, e.g. Gaussianity, that are unrealistic for most problems involving image data. More recently, some interesting ideas have been advanced under the principle that feature selection and classifier design should be solved concurrently [15, 16]. While traditionally viewed as problematic, due to the fact that the space in which training takes place becomes high-dimensional, this approach has been made feasible by powerful learning techniques, such as boosting or support vector machines, that are quite insensitive to the curse of dimensionality. Nevertheless, there are still some significant limitations, such as the fact that these techniques do not scale well in the number of image classes, or the fact that they lead to highly intensive training.

In the speech and learning communities, various authors have advocated the use of information theoretic measures for feature extraction or selection [10, 1]. These can be seen as instantiations of the *the infomax principle* of neural organization[1] proposed by Linsker [7], which also encompasses information theoretic approaches for independent compo-

[1]Under the infomax principle, the optimal organization for a complex multi-layered perceptual system is one where the information that reaches each layer is processed so that the maximum amount of information is preserved for subsequent layers.

nent analysis and blind-source separation [3]. In the classification context, infomax recommends the selection of the feature transform that maximizes the mutual information (MI) between features and class labels. While searching for the features that preserve the maximum amount of information about the class is, at an intuitive level, an appealing discriminant criteria, the infomax principle does not establish a direct connection to the ultimate measure of classification performance - the *probability of error* (PE). An alternative, that has received some attention in the speech literature [11], is to minimize Bayes error (BE), the tightest possible classifier-independent lower-bound on the PE.

We have recently shown that the two strategies (infomax and minimum BE) are very closely related. In particular, infomax has been shown to be equivalent to the minimization of a lower bound on BE that is tight and whose extrema are co-located with those of the BE. It follows from these results that *infomax solutions are very close to optimal in the minimum BE sense*, providing a formal justification for the use of infomax as a discriminant criteria. The analysis of some simple classification problems also revealed that a quantity which plays an important role in infomax solutions is the marginal diversity: the average distance between each of the marginal class-conditional densities and their mean. This inspired a generic principle for feature selection, *the principle of maximum marginal diversity* (MMD), that only requires marginal density estimates and can therefore be implemented with extreme computational simplicity.

In this paper, after reviewing these results, we characterize the problems for which the MMD principle is guaranteed to be optimal in the infomax sense. We derive a set of sufficient conditions for the equality of marginal diversity and the infomax cost and study their implications for visual recognition. These conditions turn out to be supported by evidence from various recent studies on the statistics of biologically plausible image transformations [9, 5]. This suggests that *in the context of visual recognition, MMD feature selection will lead to solutions that are optimal in the infomax sense*. Given the computational simplicity of the MMD principle, this is quite significant. We present the results of various experiments that 1) demonstrate the superiority of marginal diversity over energy compaction as a cost function for feature selection, and 2) provide evidence that features extracted through MMD can correlate well with those deemed as perceptually relevant for low-level image classification.

2 Infomax vs minimum Bayes error

We start by reviewing the relationships between the infomax cost and BE.

Theorem 1 *Given a classification problem with M classes in a feature space $\mathcal{X}$, the decision function which minimizes the probability of classification error is the Bayes classifier $g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})$, where Y is a random variable that assigns $\mathbf{x}$ to one of M classes, and $i \in \{1, \ldots, M\}$. Furthermore, the PE is lower bounded by the* Bayes error

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{1}$$

where $E_{\mathbf{x}}$ means expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$.

Proof: see the appendix for all proofs.

Principle 1 (infomax) *Consider an M-class classification problem with observations drawn from random variable $\mathbf{Z} \in \mathcal{Z}$, and the set of feature transformations $T : \mathcal{Z} \to \mathcal{X}$. The best feature space is the one that maximizes the mutual information $I(Y; \mathbf{X})$ where Y is the class indicator variable defined above, $\mathbf{X} = T(\mathbf{Z})$, and $I(Y; \mathbf{X}) = \sum_i \int p_{\mathbf{X},Y}(\mathbf{x}, i) \log \frac{p_{\mathbf{X},Y}(\mathbf{x},i)}{p_{\mathbf{X}}(\mathbf{x})p_Y(i)} d\mathbf{x}$ the mutual information between $\mathbf{X}$ and Y.*

It is straightforward to show that $I(\mathbf{X}; Y) = H(Y) - H(Y|\mathbf{X})$, where $H(\mathbf{X}) = -\int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$ is the entropy of $\mathbf{X}$. Since the class entropy $H(Y)$ does not depend on T, infomax is equivalent to the minimization of the the posterior entropy $H(Y|\mathbf{X})$. Combining this with the following result shows that *infomax minimizes a lower bound on BE*.

Theorem 2 *The BE of an M-class classification problem with feature space $\mathcal{X}$ and class indicator variable Y, is lower bounded by*

$$L^*_{\mathcal{X}}(M) \geq \frac{1}{\log M} H(Y|\mathbf{X}) - \frac{\log(2M-1)}{\log M} + 1, \tag{2}$$

where $\mathbf{X} \in \mathcal{X}$ is the random vector from which features are drawn. When M is large ($M \to \infty$) this bound reduces to $L^*_{\mathcal{X}}(M) \geq \frac{1}{\log M} H(Y|\mathbf{X})$.

In fact, the proof of this theorem shows that the LHS of (2) is a good approximation to its RHS. It follows that *infomax solutions will, in general, be very similar to those that minimize the BE* . We omit the details here (see [13] for a complete derivation) but illustrate this fact by a simple example in Figure 1. Notice that the approximation is particularly good for optimization purposes since the extrema of the two functions are co-located.

2.1 Feature selection

Because the possible number of feature subsets in a feature selection problem is combinatorial, feature selection techniques rely on *sequential search* methods [6]. These methods proceed in a sequence of steps, each adding a set of

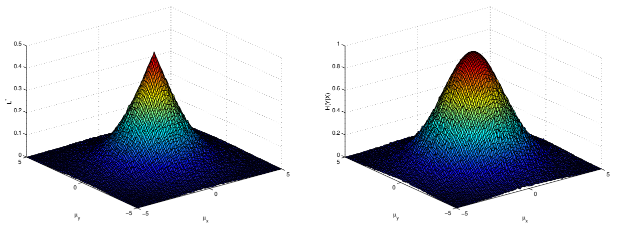

Figure 1: The LHS of (2) as an approximation to (1) for a two-class Gaussian problem where $P_{\mathbf{X}|Y}(\mathbf{x}|1) \sim N(\mathbf{0}, \mathbf{I})$ and $P_{\mathbf{X}|Y}(\mathbf{x}|2) \sim N(\mu, \mathbf{I})$. All plots are functions of μ. Left: surface plot of (1). Right: surface plot of the LHS of (2).

features to the current best subset, with the goal of optimizing a given cost function[2]. We denote the current subset by $\mathbf{X}_c$, the added features by $\mathbf{X}_a$ and the new subset by $\mathbf{X}_n = (\mathbf{X}_a, \mathbf{X}_c)$.

Theorem 3 *Consider an M-class classification problem with observations drawn from a random variable $\mathbf{Z} \in \mathcal{Z}$, and a feature transformation $T : \mathcal{Z} \to \mathcal{X}$. $\mathcal{X}$ is a infomax feature space if and only if $\forall T' \neq T$*

$$\left\langle KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)||P_{\mathbf{X}}(\mathbf{x})\right]\right\rangle_Y \geq \left\langle KL\left[P_{\mathbf{X}'|Y}(\mathbf{x}|i)||P_{\mathbf{X}'}(\mathbf{x})\right]\right\rangle_Y \quad (3)$$

where $\mathbf{X} = T(\mathbf{Z})$, $\mathbf{X}' = T'(\mathbf{Z})$, $\langle f(i)\rangle_Y = \sum_i P_Y(i)f(i)$ denotes expectation with respect to the prior class probabilities and $KL[p||q] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ is the Kullback-Leibler divergence between p and q. Furthermore, if $\mathbf{X}_n = (\mathbf{X}_a, \mathbf{X}_c)$, the infomax cost function decouples into two terms according to

$$\begin{aligned} &\left\langle KL\left[P_{\mathbf{X}_n|Y}(\mathbf{x}_n|i)||P_{\mathbf{X}_n}(\mathbf{x}_n)\right]\right\rangle_Y = \\ &= \left\langle KL\left[P_{\mathbf{X}_a|\mathbf{X}_c,Y}(\mathbf{x}_a|\mathbf{x}_c,i)||P_{\mathbf{X}_a|\mathbf{X}_c}(\mathbf{x}_a|\mathbf{x}_c)\right]\right\rangle_Y \\ &+ \left\langle KL\left[P_{\mathbf{X}_c|Y}(\mathbf{x}_c|i)||P_{\mathbf{X}_c}(\mathbf{x}_c)\right]\right\rangle_Y . \end{aligned} \quad (4)$$

Equation (3) exposes the discriminant nature of the infomax criteria. Noting that $P_{\mathbf{X}}(\mathbf{x}) =< P_{\mathbf{X}|Y}(\mathbf{x}|i) >_Y$, it clearly favors feature spaces where each class-conditional density is as distant as possible (in the KL sense) from the average among all classes. This is a sensible way to quantify the intuition that optimal discriminant transforms are the ones that best separate the different classes. Equation (4), in turn, leads to an optimal rule for finding the features $\mathbf{X}_a$ to merge with the current optimal solution $\mathbf{X}_c$: the set which minimizes $\left\langle KL\left[P_{\mathbf{X}_a|\mathbf{X}_c,Y}(\mathbf{x}_a|\mathbf{x}_c,i)||P_{\mathbf{X}_a|\mathbf{X}_c}(\mathbf{x}_a|\mathbf{x}_c)\right]\right\rangle_Y$.

[2]These methods are called *forward search* techniques. There is also an alternative set of *backward search* techniques, where features are successively removed from an initial set containing all features. We ignore the latter for simplicity, even though all that is said can be applied to them as well.

These observations have motivated the introduction of the principle of feature selection by maximum marginal diversity in [13].

2.2 Maximum marginal diversity

The simplest solution to (4) is to consider the situation in which the set of added features $\mathbf{X}_a$ has cardinality one, i.e. features are added one at a time. This leads to the notion of marginal diversity,

Definition 1 *Consider a classification problem on a feature space $\mathcal{X}$, and a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ from which feature vectors are drawn. Then, $\mathbf{md}(X_k) =< KL[P_{X_k|Y}(x|i)||P_{X_k}(x)] >_Y$ is the marginal diversity of feature X_k.*

Comparing to (3), it is clear that MD is equivalent to the infomax cost for one-dimensional problems, i.e. the selection of the best feature. For higher dimensional problems, the principle of maximum marginal diversity advocates the approximation of the infomax cost by the sum of marginal diversities.

Principle 2 (Maximum marginal diversity) *The best solution for a feature selection problem is to select the subset of features that leads to a set of maximally diverse marginal densities.*

A significant advantage of optimizing MD rather than the infomax cost (3) is computational. In fact, it is straightforward to implement the MMD principle with the following algorithm.

Algorithm 1 (MMD feature selection) *For a classification problem with n features $\mathbf{X} = (X_1, \ldots, X_n)$, M classes $Y \in \{1, \ldots, M\}$ and class priors $P_Y(i) = p_i$ the following procedure returns the top N MMD features.*

- *foreach feature $k \in \{1, \ldots, n\}$:*
  - * *foreach class $i \in \{1, \ldots, M\}$, compute an histogram estimate $\mathbf{h}_{k,i}$ of $P_{X_k|Y}(x|i)$,*
  - * *compute $\mathbf{h}_k = \frac{1}{M}\sum_i h_{k,i}$,*
  - * *compute the marginal diversity $\mathbf{md}(X_k) = \sum_i p_i \mathbf{h}_{k,i}^T \log(\mathbf{h}_{k,i}./\mathbf{h}_k)$, where both the log and division ./ are performed element-wise,*
- *order the features by decreasing diversity, i.e. find $\{k_1, \ldots, k_n\}$ such that $\mathbf{md}(X_{k_i}) \geq \mathbf{md}(X_{k_{i+1}})$*
- *return $\{X_{k_1}, \ldots, X_{k_N}\}$.*

An important point is that simplicity is not achieved by relaxing the requirement that a good feature transformation should be inherently discriminant. By recommending the elimination of the dimensions along which the projections of the class densities are most similar, MMD clearly

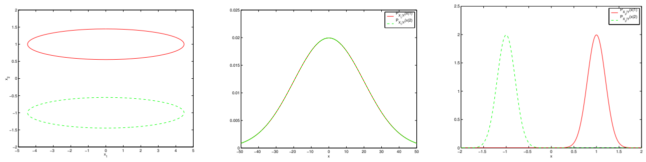

Figure 2: Gaussian problem with two classes $Y \in \{1,2\}$, in the two-dimensions, $\mathbf{X} = (X_1, X_2)$. Left: contours of 65% probability. Middle: marginals for X_1. Right: marginals for X_2. In this problem, PCA selects X_1 as the most important feature, despite the fact that the marginals $P_{X_1|Y}(x|1)$ and $P_{X_1|Y}(x|2)$ are equal and feature X_1 does not contain any useful information for classification. The MMD principle correctly selects feature X_2.

satisfies this requirement. This is unlike feature extraction/selection methods based on principles such as decorrelation or energy compaction, e.g. principal component analysis (PCA), that are prevalent in the vision literature. It is not hard to find examples where the latter can lead to the worst possible solution while the former is still able to reach optimal decisions (see Figure 2 for one such example).

3 MMD vs Infomax

While, as seen above, MMD feature selection is guaranteed to find the optimal infomax solution for one-dimensional problems, no such guarantees exist in higher dimensions. In this section we seek a precise understanding of the relationships between MMD and infomax. These relationships are summarized by the following result.

Theorem 4 *Consider a classification problem with class labels drawn from a random variable Y and features drawn from a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ and let $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ be the optimal feature subset of size N in the infomax sense. Then*

$$\left\langle KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)||P_{\mathbf{X}}(\mathbf{x})\right]\right\rangle_Y = \sum_{k=1}^{N} \mathbf{md}(X_k^*) + \\ + \sum_{k=2}^{N} I(X_k^*; \mathbf{X}_{1,k-1}^*|Y) - \sum_{k=2}^{N} I(X_k^*; \mathbf{X}_{1,k-1}^*) \quad (5)$$

where $\mathbf{X}_{1,k-1}^* = \{X_1^*, \ldots, X_{k-1}^*\}$

Equation (5) establishes three requirements for the optimal set of features: they must have 1) large marginal diversity, 2) low mutual information, and 3) high mutual information given the image class. These requirements can seen as enforcing three very intuitive principles.

1. *Discrimination*: each feature in the optimal set must be discriminant.
2. *Feature diversity*: the features in the optimal set must not be redundant.
3. *Reinforcement* : the only important dependencies between features are those that carry information about the class Y.

Notice that feature diversity is different from marginal diversity, the driving principle for MMD. While the latter guarantees discrimination, it does not necessarily lead to a compact code. For example, two features that are exact replicas of each other will exhibit the same marginal diversity and will therefore be simultaneously selected or rejected by MMD. This is not desirable, since the selection of two features that are equivalent represents a waste of the available space dimensions. On the other hand, simply penalizing mutual dependencies is overkill since such dependencies may be crucial for fine discrimination between otherwise similar classes. In this sense, the third principle guarantees that the whole is more than the sum of the parts.

While this three-factor decomposition of the infomax cost is conceptually interesting, it suffers from the practical limitation that it is usually difficult to evaluate mutual information in high dimensions. It is therefore not clear that, in practice, relying on the infomax cost will guarantee better solutions than those made available by MMD. An alternative path of interest is to seek a precise characterization of the problems where MMD is indeed equivalent to infomax. Such characterization is provided by the following Corollary of the theorem above.

Corollary 1 *Consider a classification problem with class labels drawn from a random variable Y and features drawn from a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ and let $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ be the optimal feature subset of size N in the infomax sense. If*

$$I(X_k^*; \mathbf{X}_{1,k-1}^*) = I(X_k^*; \mathbf{X}_{1,k-1}^*|Y), \forall k \in \{1, \ldots, N\} \quad (6)$$

where $\mathbf{X}_{1,k-1}^* = \{X_1^*, \ldots, X_{k-1}^*\}$, the set $\mathbf{X}^*$ is also the optimal subset of size N in the MMD sense. Furthermore,

$$\left\langle KL\left[P_{\mathbf{X}^*|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*}(\mathbf{x})\right]\right\rangle_Y = \sum_{k=1}^{N} \mathbf{md}(X_k^*). \quad (7)$$

The corollary states that the MMD and infomax solutions are identical when the mutual information between features is not affected by knowledge of the class label. Or, in other words, when the dependence between features is not class-dependent. This is an interesting condition in light of various recent studies that have reported the observationof consistent patterns of dependence between the features of various biologically plausible image transformations [9, 5]. For example, spatially co-located wavelet coefficients at adjacent scales tend to be dependent, exhibiting

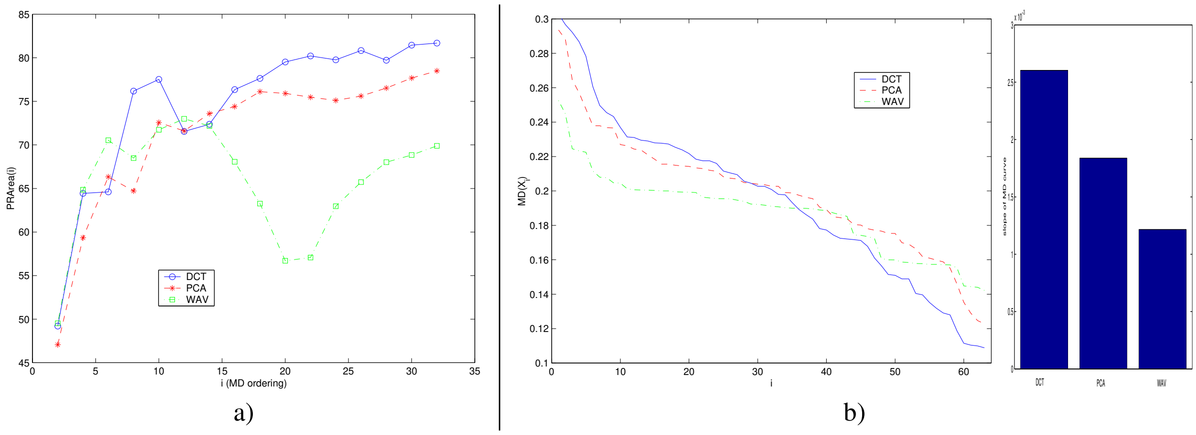

Figure 3: a) Curves of PRA as a function of the number of features used for retrieval. b) Curves of MD as a function of the same number (left) and magnitude of the slope of the MD curve for each of the feature transformations considered (right).

the same pattern dependence (bow-shaped conditional densities) across a wide variety of imagery [9]. Even though the fine details of feature dependence may vary from one image class to the next, these studies suggest that the coarse structure of the patterns of dependence between such features follow universal statistical laws that hold for all types of images. The potential implications of this conjecture are quite significant. First it implies that, in the context of visual processing, (6) will be approximately true and the MMD principle will consequently lead to solutions that are very close to optimal, in the minimum BE sense. Given the simplicity of MMD feature selection, this is quite remarkable. Second, it implies that when combined with such transformations, the marginal diversity is a close predictor for the infomax cost (and consequently the BE) achievable in a given feature space. This enables quantifying the goodness of the transformation without even having to build the classifier, and leading to further computational simplicity.

4 Experimental results

In this section we report on experiments designed to quantify the effectiveness of MMD as a feature selection technique for visual recognition. These experiments were conducted in the context of texture classification and retrieval using the Brodatz database[3]. The experimental set up is basically the same as that reported in some of our previous work (e.g. see [14]) and the reader is referred to those references for more details. In a nutshell, the database is divided into a training and test set, the training set used for all the learning and the test set for evaluation. There are a total of 112 image classes, each containing 9 images, 8 of which are used for training. Features are extracted from random 8×8 image neighborhoods and all classification/retrieval results are obtained with classifiers based on Gaussian mixtures. Three feature transformations were considered, building on previous experience that space/space-frequency transforms tend to work well for this task [14]: the discrete cosine transform (DCT), a wavelet representation (WAV), and principal component analysis (PCA).

4.1 MD as a predictor of feature goodness

The first experiment was designed to evaluate how the MMD rankings relate to the actual retrieval/recognition performance of various feature sets. For this, we designed a retrieval experiment where the images in the training set were considered as a visual database, and the images in the test set as a set of visual queries. A set of $1,000$ feature vectors was extracted from each of the images in the training set to create a sample with $8,000$ vectors per class. The resulting 112 samples were then fed to the MMD algorithm in order to find the most discriminant features for the retrieval problem. The process was repeated for each of the three feature transformations and, for each subspace dimension, the query images ranked according to their class-posterior probabilities. Precision/recall (PR) curves were then measured for all subspace dimensions. To simplify the presentation, we summarize each PR curve by its integral, i.e. the area under the PR curve, which we denote by precision/recall area (PRA).

Figure 3 a) presents the curve of PRA, as a function of the number of features used in the retrieval operation, for the three feature transformations. It is safe to conclude that the DCT achieves the best performance, followed by PCA, and that the wavelet is the worst performer. Figure 3

[3]We are currently performing the same type of experiments for object recognition on the COIL database, and image retrieval on Corel. The results of the complete evaluation will be included in the final paper.

b), presents the equivalent curves for the MD, i.e.the MD as a function of subspace dimension, as well as the absolute value of the slope of the line that best fits (in the least squares sense) each of the MD curves. It is clear that ranking the MD curves by the magnitude of this slope leads to a ranking identical to that obtained by the actual measurement of the PRA curves. This makes intuitive sense since, while a large slope indicates that the features range from highly discriminant to poorly discriminant, a small slope is indicative of an an homogeneous set of features that are all equally discriminant. Therefore, when only a small percentage of the features is used, the retained subset will be more discriminant for transformations of the former type. Or, in other words, MD curves of larger slope indicate more compaction of the discriminant power into a small subset of the features.

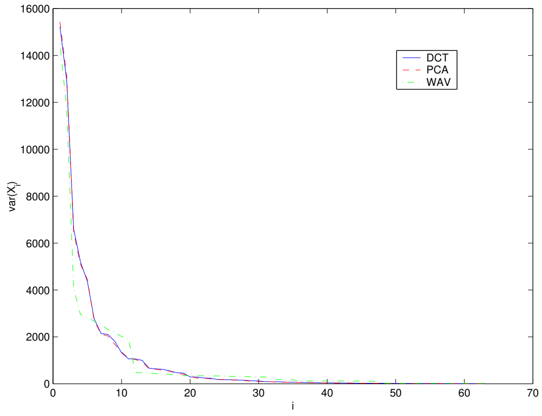

Figure 4: Curve of feature variance as a function of subspace dimension (features ordered by decreasing variance).

For completeness, we also show in Figure 4 the curves of feature variance as a function of subspace dimension (when the features are ordered by decreasing variance). Note that, unlike MD, it is quite hard to infer from these plots which transformation will lead to better recognition. In fact, the variance curves for the DCT and PCA transforms are virtually indistinguishable! This observation (and the fact that PCA, which is theoretically optimal from an energy compaction point of view, performs worse than the DCT!), are further evidence that energy compaction is a poor metric for feature selection.

4.2 Perceptual relevance

The second experiment was designed to evaluate if the features extracted by MMD would correlate with image properties that are deemed perceptually relevant. Once again, we extracted $8,000$ feature vectors from each class, but now the 112 samples were combined in order to find the features that best discriminate between a given class (the target class) and everything else in the database. For this, we grouped all the feature vectors from classes other than the target in a large sample. This and the sample from the target class were then fed to the MMD algorithm. The procedure was repeated for all classes in the database, each taking its turn as the target class. The goal was to identify the most discriminant features for each image class, and evaluate if they correlate with perceptually salient features of the images in that class. Note that the objective was not to make sweeping claims about the role of MMD in human perception, but simply to investigate if MMD could extract meaningful features from classes that contain local visual attributes of obvious perceptual relevance (e.g. bars, lines, corners, and so forth). We believe that the ability to do so is an important step towards recognition/retrieval systems capable of image similarity judgments that mimic those made by humans.

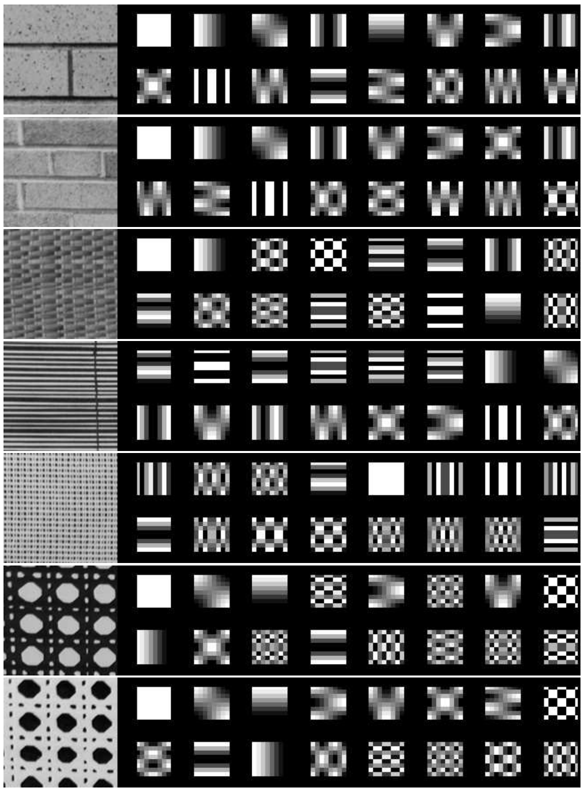

Figure 5: Representative images from 7 texture classes on Brodatz (left) and the corresponding 16 most discriminating features as determined by the MMD algorithm (right).

As Figures 5 and 6 illustrate, the results were highly encouraging. Each picture in Figure 5 presents a representative image from a texture class (on the right) and blown-up replicas of the 16 top features selected by MMD as most discriminant for that class (on the left). For calibration, the images on the left have size 128×128 while each of the features is of size 8×8. One interesting observation that can be made from the figure is that the extracted sets of features are stable with respect to perturbations such as scaling or figure-ground reversal. In particular, even though there is a variation in scale of about 2 to 1 between the top two images (and the texture is not even exactly the same!) the corresponding sets of optimal features share 7 out of the 8 top elements and 14 out of the top 16. Similarly the optimal sets for the last two images share 7 of the top 8 and 13 of the top 16. This stability is interesting in the sense that it indicates invariance of the representation to non-trivial image transformations. The question of invariance is one that we intend to study more systematically in future work.

A second interesting observation is that the perceptually most salient aspects of each texture class seem to be covered by the extracted features. In particular, 1) the optimal set for the two top images includes corner, line, and t-junction detectors and all of the these appear high up in the list (after the mean and horizontal gradient); 2) the optimal features for the third image are similar to those of the first two, but of higher frequency; 3) image four originates detectors of horizontal lines that are closely spaced vertically; and 4) the fifth image results in detectors of high-frequency patterns. This is also illustrated by Figure 6 where we show the image responses of the 5 most discriminant features (other than the mean) for the two image classes at the top of Figure 5. Once again, it is visible that the features are basically detectors for the presence of bars, corners and t-junctions.

Appendix
For the proofs of Theorems 1-3 see [13].

A Proof of Theorem 4

Applying (4), with $\mathbf{X}_a = X_N^*$ and $\mathbf{X}_c = \mathbf{X}^*_{1,N-1}$ leads to

$$\begin{aligned}
&\left\langle KL\left[P_{\mathbf{X}^*|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*}(\mathbf{x})\right]\right\rangle_Y = \\
&= \left\langle KL\left[P_{X_N^*|\mathbf{X}^*_{1,N-1},Y}(x|\mathbf{x},i)||P_{X_N^*|\mathbf{X}^*_{1,N-1}}(x|\mathbf{x})\right]\right\rangle_Y \\
&+ \left\langle KL\left[P_{\mathbf{X}^*_{1,N-1}|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*_{1,N-1}}(\mathbf{x})\right]\right\rangle_Y,
\end{aligned}$$

and, by repeating this procedure recursively,

$$\begin{aligned}
&\left\langle KL\left[P_{\mathbf{X}^*|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*}(\mathbf{x})\right]\right\rangle_Y = \\
&= \sum_{k=2}^{N}\left\langle KL\left[P_{X_k^*|\mathbf{X}^*_{1,k-1},Y}(x|\mathbf{x},i)||P_{X_k^*|\mathbf{X}^*_{1,k-1}}(x|\mathbf{x})\right]\right\rangle_Y \\
&+ \left\langle KL\left[P_{X_1^*|Y}(x|i)||P_{X_1^*}(x)\right]\right\rangle_Y
\end{aligned}$$

Noting that

$$\begin{aligned}
&\left\langle KL\left[P_{\mathbf{X}_a|\mathbf{X}_c,Y}(\mathbf{x}_a|\mathbf{x}_c,i)||P_{\mathbf{X}_a|\mathbf{X}_c}(\mathbf{x}_a|\mathbf{x}_c)\right]\right\rangle_Y = \\
&= \sum_i \int P_{\mathbf{X}_a,\mathbf{X}_c,Y}(\mathbf{x}_a,\mathbf{x}_c,i)\log\frac{P_{\mathbf{X}_a|\mathbf{X}_c,Y}(\mathbf{x}_a|\mathbf{x}_c,i)}{P_{\mathbf{X}_a|Y}(\mathbf{x}_a|i)} \\
&+ \sum_i \int P_{\mathbf{X}_a,Y}(\mathbf{x}_a,i)\log\frac{P_{\mathbf{X}_a|Y}(\mathbf{x}_a|i)}{P_{\mathbf{X}_a}(\mathbf{x}_a)} \\
&+ \int P_{\mathbf{X}_a,\mathbf{X}_c}(\mathbf{x}_a,\mathbf{x}_c)\log\frac{P_{\mathbf{X}_a}(\mathbf{x}_a)}{P_{\mathbf{X}_a|\mathbf{X}_c}(\mathbf{x}_a|\mathbf{x}_c)} \\
&= \left\langle KL\left[P_{\mathbf{X}_a|Y}(\mathbf{x}_a|i)||P_{\mathbf{X}_a}(\mathbf{x}_a)\right]\right\rangle_Y \\
&- \left[I(\mathbf{X}_a;\mathbf{X}_c) - I(\mathbf{X}_a;\mathbf{X}_c|Y)\right]
\end{aligned}$$

letting $\mathbf{X}_a = X_k^*$ and and $\mathbf{X}_c = \mathbf{X}^*_{1,k-1}$, this equality can also be written as

$$\begin{aligned}
&\left\langle KL\left[P_{\mathbf{X}^*|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*}(\mathbf{x})\right]\right\rangle_Y = \\
&= \sum_{k=1}^{N}\left\langle KL\left[P_{X_k^*|,Y}(x|i)||P_{X_k^*}(x)\right]\right\rangle_Y \\
&- \sum_{k=2}^{N}\left[I(X_k^*;\mathbf{X}^*_{1,k-1}) - I(X_k^*;\mathbf{X}^*_{1,k-1}|Y)\right]
\end{aligned}$$

and the theorem follows.

B Proof of Corollary 1

From Theorem 4, if (6) holds,

$$\begin{aligned}
&\left\langle KL\left[P_{\mathbf{X}^*|Y}(\mathbf{x}|i)||P_{\mathbf{X}^*}(\mathbf{x})\right]\right\rangle_Y = \\
&= \sum_{k=1}^{N}\mathbf{md}(X_k^*) \\
&= \sum_{k=1}^{N}\left\langle KL\left[P_{X_k^*|,Y}(x|i)||P_{X_k^*}(x)\right]\right\rangle_Y.
\end{aligned}$$

Since, from the properties of the KL divergence [4], the terms $\left\langle KL\left[P_{X|,Y}(x|i)||P_X(x)\right]\right\rangle_Y$ are always non-negative, the sum is maximized when each of the terms is maximum. It follows that the optimal infomax features are the ones with maximum marginal diversity and infomax is therefore equivalent to MMD.

References

[1] S. Basu, C. Micchelli, and P. Olsen. Maximum Entropy and Maximum Likelihood Criteria for Feature Selection from Multivariate Data. In *Proc. IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland,2000.

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

[3] A. Bell and T. Sejnowski. An Information Maximisation Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[5] J. Huang and D. Mumford. Statistics of Natural Images and Models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, 1999.

[6] A. Jain and D. Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[7] R. Linsker. Self-Organization in a Perceptual Network. *IEEE Computer*, 21(3):105–117, March 1988.

[8] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearence. *International Journal of Computer Vision*, 14:5–24, 1995.

[9] J. Portilla and E. Simoncelli. Texture Modeling and Synthesis using Joint Statistics of Complex Wavelet Coefficients. In *IEEE Workshop on Statistical and Computational Theories of Vision, Fort Collins, Colorado*, 1999.

[10] J. Principe, D. Xu, and J. Fisher. Information-Theoretic Learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering, Volume 1: Blind-Source Separation*. Wiley, 2000.

[11] G. Saon and M. Padmanabhan. Minimum Bayes Error Feature Selection for Continuous Speech Recognition. In *Proc. Neural Information Proc. Systems*, Denver, USA, 2000.

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

[13] N. Vasconcelos. Feature Selection by Maximum Marginal Diversity. In *Neural Information Processing Systems, Vancouver, Canada*, 2002.

[14] N. Vasconcelos and G. Carneiro. What is the Role of Independence for Visual Regognition? In *Proc. European Conference on Computer Vision, Copenhagen, Denmark*, 2002.

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

[16] M. Yang, D. Roth, and N. Ahuja. Learning to Recognize 3D Objects with SNoW. In *Proc.* 6^{th} *European Conference on Computer Vision, Dublin, Ireland*, 2000.

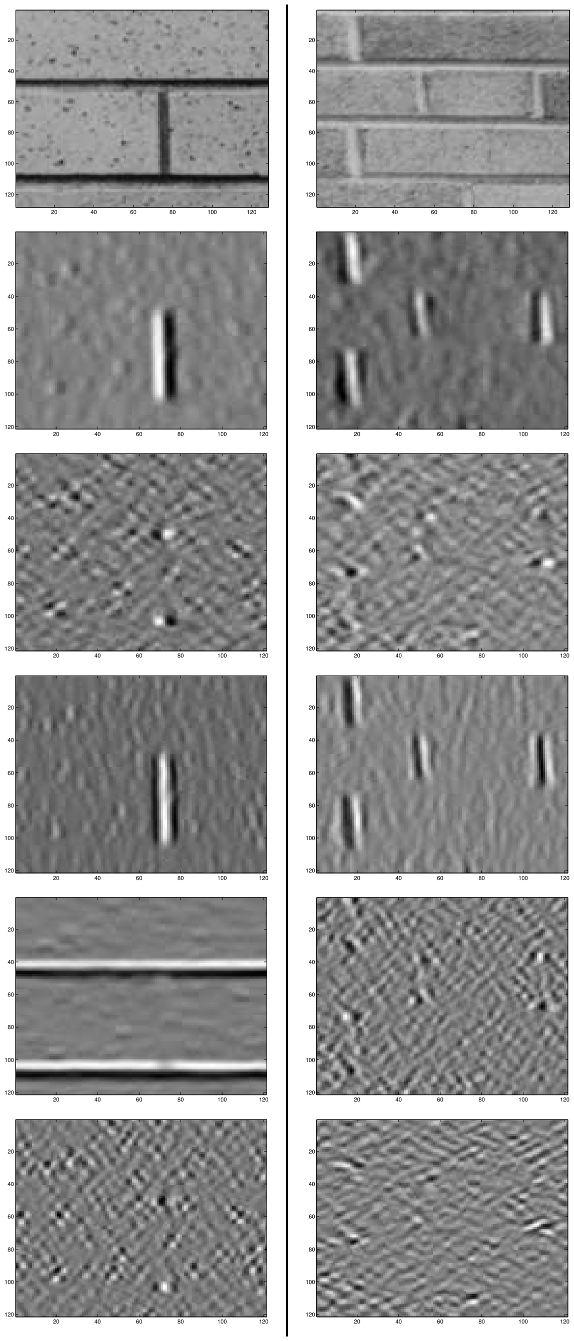

Figure 6: Two textures from Brodatz (top) and the corresponding responses of the 5 most discriminant features for the class of each image.