Shorter version appears in IEEE Conf. on Computer Vision and Pattern Recognition, Washington DC, 2004.

Scalable Discriminant Feature Selection for Image Retrieval and Recognition

Nuno Vasconcelos
Statistical Visual Computing Laboratory,
University of California San Diego
nuno@ece.ucsd.edu

Manuela Vasconcelos
Harvard Robotics Laboratory
Harvard University
mana@hrl.harvard.du

Abstract

Problems such as object recognition or image retrieval require feature selection (FS) algorithms that scale well enough to be applicable to databases containing large numbers of image classes and large amounts of data per class. We exploit recent connections between information theoretic feature selection and minimum Bayes error solutions to derive FS algorithms that are optimal in a discriminant sense without compromising scalability. We start by formalizing the intuition that optimal FS must favor discriminant features while penalizing discriminant features that are redundant. We then rely on this result to derive a new family of FS algorithms that enables an explicit trade-off between complexity and classification optimality. This trade-off is controlled by a parameter that encodes the order of feature redundancies that must be explicitly modeled to achieve the optimal solution. Experimental results on databases of natural images show that this order is usually low, enabling optimal FS with very low complexity. In particular, the algorithms now proposed are shown to significantly outperform currently available scalable alternatives: from purely discriminant approaches to approaches that only emphasize redundancy-reduction, e.g. the commonly used variance compaction methods.

1. Introduction

Various challenging problems in vision, including visual recognition, image retrieval, or the recognition of people and events, can be formalized as statistical classification problems. It is well known that one crucial ingredient for success in these problems is a representation based on a discriminant set of features and, for this reason, feature selection has a long history in the machine learning and pattern recognition literatures. However, various attributes of traditional feature selection (FS) solutions, such as a significant emphasis on binary classification problems, the expectation of relatively small amounts of data, or the assumption of parametric sources, are not realistic in the vision context. In fact, for problems such as image retrieval, where 1) there are usually large numbers of visual classes to be processed, 2) the data is non-Gaussian and non-homogeneous (and the assumption of any parametric model is, therefore, highly restrictive) and 3) there is a need to process very large training sets, traditional feature selection strategies either 1) simply fail to achieve meaningful results, 2) take an unrealistic (and practically infeasible) time to compute, or 3) both[1].

Some of these limitations, such as the dependence on particular probabilistic models, have been eliminated by recent advances in machine learning and already translated into success stories for vision, e.g. the boosted face detector by Viola and Jones[2], that can be seen as a high-performance feature selection algorithm. On the other hand, these solutions exacerbate some of the limitations enumerated above, namely the unavailability of (practical) extensions to problems with more than two classes, and an immense training complexity. Due to this inherent lack of scalability, most existing FS techniques are not applicable to large-scale problems, such as retrieval or recognition.

Somewhat surprisingly, the problem of provably optimal feature design with low complexity has not yet been the subject of extensive research in either the vision or learning literatures. Consequently, in most domains where scalability is a requirement of paramount importance, feature optimality is commonly traded for computational tractability. For example, while various intuitive justifications have been offered for most feature sets commonly adopted in the retrieval literature - e.g. the biological plausibility of Gabor or wavelet representations [3, 4, 5], or the intuitive appeal of perceptually salient attributes such as color and edginess [6, 7, 8, 9, 10, 11] - little has been shown in terms of their optimality. To compound the problem, (and, once again, due to a lack of computationally feasible alternatives) it is still the norm for large scale recognition or retrieval systems to resort to sub-optimal principles, such as energy (or variance) compaction, to select the best subsets among these features [12, 6, 13, 14].

In order to avoid what appears to be an intrinsic lack of scalability of optimal feature design, when optimality is ex-

[1] For example, [1] reports that the application of a traditional feature selection technique to an image retrieval problem of relatively small dimensions (two classes, 700 examples per class) required 12 days of processing time.

plicitly defined in the minimum probability of error (MPE) sense, we have been recently pursuing alternative, and more *scalable*, formulations. In particular, we have shown that information theoretic solutions based on the maximization of the mutual information between features and class labels (the *infomax* criteria [15]) have various appealing properties for FS, in the context of vision problems [16, 17]. These properties, which are reviewed in section 2, include 1) near optimality in the MPE sense, 2) significantly better scalability than most existing FS techniques, and 3) the existence of a set of conditions under which the implementation complexity is equivalent to that of the simplest known suboptimal methods (e.g. principal component analysis, PCA) without compromise of MPE optimality.

The third property is particularly interesting because the validity of these conditions appears to be supported by various recent studies in image statistics [18, 19]. While this empirical observation is impossible to prove analytically, it is possible to test it through the following indirect empirical strategy: design FS algorithms which assume the conditions to hold and evaluate their classification optimality. This strategy motivated a FS technique based on the maximization of the marginal diversity of the selected features [16] (which consists of selecting the set of features with least overlapping marginal class-conditional distributions). While inherently discriminant, and computationally trivial, the MMD solution is sub-optimal in the presence of feature dependencies that strongly convey information about the class label[2]. This can be problematic, since our experiments indicate that, for natural images, such dependencies are not always neglectable.

In this paper we show that, when this is the case, *infomax-optimal FS* cannot be achieved by stressing discrimination alone but *requires a good balance between maximizing the discriminant power of the selected features and minimizing their redundancy*: while a highly discriminant feature can be a major asset for classification, a second highly discriminant feature that does not add much information about the class label (e.g. which simply replicates the information contained in the first) is highly undesirable[3]. In this context, MMD and variance-based techniques, such as PCA, occupy the two ends of the image representation spectrum: while the former places all emphasis on discrimination, the latter only strives to eliminate redundancies. It seems natural to expect that better performance can be attained by FS solutions that jointly address the two components of the problem. One of the interesting aspects of the information theoretic formulation is precisely that it provides the appropriate mathematical formalism for capturing all the subtleties involved, e.g. that *while generic dependencies are irrelevant, dependencies that convey information about the class label are important*.

In particular, it is shown that solutions which jointly address discrimination and redundancy reduction do exist within the infomax framework. For this, we introduce the concept of an l-decomposable set of features, i.e. a feature set that can be divided into mutually exclusive subsets of l features such that:

- features within each subset are arbitrarily dependent,
- the dependence between subsets does not convey information on the image class.

We then show that, when l-decomposability holds, infomax-optimal FS only requires density estimates of order $l+1$. For low values of the decomposability order l, this implies that optimal FS is achievable with reduced computational complexity. We next introduce a family of algorithms, parameterized by l, that allow the explicit control of the trade-off between computational complexity and the ability to model arbitrary feature dependencies. This enables us to evaluate the value of l that achieves the optimal balance between the maximization of marginal diversity and the minimization of feature redundancies. Experiments on various image databases consistently indicate that, for various of the feature transformations in common use in the retrieval and recognition literatures, optimal performance can be achieved with models that explain arbitrary dependencies of second order. The resulting algorithms are shown to substantially outperform FS based on either variance compaction or MMD alone, without significant cost in terms of computational efficiency.

2. Information theoretic feature selection

Information theory provides a principled way to capture the intuition that the best feature space for a given classification problem is the one which keeps most information about the class labels. More formally, given a M-ary classification problem with observations drawn from random variable $\mathbf{Z} \in \mathcal{Z}$, and a set of feature transformations of the form $T : \mathcal{Z} \rightarrow \mathcal{X}$, the best feature space is the one that maximizes the mutual information $I(Y;\mathbf{X})$ where Y is the class indicator variable (i.e. a random variable that takes values in $\{1,\ldots,M\}$), $\mathbf{X} = T(\mathbf{Z})$, and

$$I(Y;\mathbf{X}) = \sum_i \int p_{\mathbf{X},Y}(\mathbf{x},i) \log \frac{p_{\mathbf{X},Y}(\mathbf{x},i)}{p_{\mathbf{X}}(\mathbf{x}) p_Y(i)} d\mathbf{x}. \quad (1)$$

Information-theoretic feature selection (ITFS) has various appealing properties that we summarize in this section (see [17] for a more detailed discussion).

[2] Note that these are distinct from generic feature dependencies, which are prevalent but do not affect the performance of MMD FS.

[3] Because it consumes a valuable resource - a slot in the set of selected features - without adding any discriminant power.

2.1. Bayes error

We start by recalling that the tightest possible lower bound on the probability of error achievable by any classifier on a given classification problem is the *Bayes error*. For an M-class problem on a feature space $\mathcal{X}$ it is given by [20]

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \quad (2)$$

where $E_{\mathbf{x}}$ means expectation with respect to $P_{\mathbf{X}}(\mathbf{x}), \mathbf{x} \in \mathcal{X}$. Since the Bayes error depends only on the selection of the feature space $\mathcal{X}$, and there is at least one classifier whose probability of error is equal to (2), the Bayes classifier, the Bayes error is the ultimate discriminant measure for FS. However, due to the nonlinearity of the $\max(\cdot)$ operator, it can be difficult to work with this cost directly. This is particularly true in the FS context, where the combinatorial complexity of finding a globally optimal solution is usually avoided by relying on greedy procedures (e.g. sequential search [21]) that, to obtain an optimal solution in high-dimensions, start from a low dimensional solution and incrementally add features to it. Since, due to the nonlinearity of (2), it is impossible to decompose the overall cost into a function of simpler terms depending only on feature subsets, these type of strategies cannot be applied to the minimization of Bayes error.

2.2. Information theoretic costs

In the FS context, mutual information is an appealing alternative to the Bayes error for two main reasons: 1) it is significantly easier manipulate, and 2) it is possible to establish formal connections between the two costs. In particular, the following properties hold [17].

Properties 1 (mutual information)

1. $I(\mathbf{X};Y) = H(Y) - H(Y|\mathbf{X})$, *where* $H(\mathbf{X}) = -\int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$ *is the entropy of* $\mathbf{X}$ *[22].*

2. $I(\mathbf{X};Y)$ *can be written as* $I(\mathbf{X};Y) = \left\langle KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)||P_{\mathbf{X}}(\mathbf{x})\right]\right\rangle_Y$ *where* $KL[p||q] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ *is the Kullback-Leibler divergence between* p *and* q*, and* $\langle f(i)\rangle_Y = \sum_i P_Y(i) f(i)$ *the expectation with respect to the class priors.*

3. *up to scaling,* $H(Y|\mathbf{X})$ *is a lower bound on the Bayes error. The bound is tight in a well-defined sense, and provides a close approximation in most situations of practical interest.*

4. *if* $\mathcal{Z}$ *is* n*-dimensional and* $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ *the optimal feature subset of size* N*, then*

$$I(\mathbf{X}^*, Y) = \sum_{k=1}^{N} I(X_k^*, Y) - \sum_{k=2}^{N} [I(X_k^*; \mathbf{X}_{1,k-1}^*) - I(X_k^*; \mathbf{X}_{1,k-1}^*|Y)]. \quad (3)$$

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

It follows from Property 1 that maximizing the mutual information is equivalent to minimizing the posterior entropy $H(Y|\mathbf{X})$ (since $H(Y)$ does not depend on the feature space). This provides an intuitive justification for ITFS, as the search for the feature space where the uncertainty about which class is responsible for each observation is minimized. Property 2 shows that the ITFS solution is equivalent to maximizing the average KL divergence (over all classes) between the class-conditional densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and the unconditional feature density $P_{\mathbf{X}}(\mathbf{x})$. Since the latter is just the average (over all classes) of the $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and the KL divergence is a measure of dissimilarity between probability densities, the property shows that ITFS is inherently discriminant: it rewards feature spaces where the density of each class is as distant as possible from the average density over all classes. Property 3 provides a formal justification for ITFS as a discriminant FS technique by establishing a connection to the Bayes error. This connection suggests that infomax solutions are also optimal in the minimum Bayes error sense. We omit the details here, see [16] for a complete presentation. Finally, Property 4 reveals two interesting properties of the information theoretic formulation of the FS problem. First, it formalizes the statement that information theoretic costs are easier to manipulate than the Bayes error, by explicitly providing a decomposition of the optimal infomax cost into a sum of simpler terms. Second, it unveils an interesting interpretation of *optimal FS as a trade-off between the maximization of discriminant power and the minimization of redundancy*.

To understand this, it helps to interpret X_k as the k^{th} most important feature, and $\mathbf{X}_{1,k-1}^*$ as the set of features that must already have been selected by the time X_k is. While, from Property 2, the first summation in (3) is a measure of the individual discriminant power of the optimal features, the second summation is a penalty on all combinations of X_k and $\mathbf{X}_{1,k-1}^*$ that *are jointly informative about the class label* Y[4]. This compensates for any double-counting of discriminant power due to the first term: when a feature is highly discriminant, all other features that are highly correlated with it are also discriminant, but would not add much to the overall discriminant power of the selected feature set. Hence, the second term penalizes *all redundancies that carry information about the class label*. Note that a direct corollary of this observation is that, for FS purposes, *all redundancy that does not carry information about the class labels can be safely ignored*. This im-

[4] Note that this term is zero when X_k and $\mathbf{X}_{1,k-1}^*$ are jointly independent of, i.e. completely uninformative about, the class label Y.

plies that independent modeling of features that are highly correlated may not necessarily lead to a loss of optimality.

2.3. Connection to image statistics

The last point is particularly interesting in light of various recent studies on image statistics, that have reported the observationof universal patterns of dependence between the features of various biologically plausible image transformations [18, 19]. For example, spatially co-located wavelet coefficients at adjacent scales tend to be dependent, exhibiting the same pattern dependence (bow-shaped conditional densities) across a wide variety of imagery [18]. 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 independently of the image class. The potential ramifications of this conjecture are quite significant since it implies that, in the context of visual processing,

$$\frac{1}{N-1}\sum_{k=2}^{N} I(X_k^*; \mathbf{X}_{1,k-1}^*) \approx \frac{1}{N-1}\sum_{k=2}^{N} I(X_k^*; \mathbf{X}_{1,k-1}^*|Y), \tag{4}$$

in which case (3) reduces to

$$I(\mathbf{X}^*, Y) \approx \sum_{k=1}^{N} I(X_k^*, Y). \tag{5}$$

The k^{th} term on this summation is referred to as the marginal diversity of feature X_k. The practical significance of this result is that the optimal solution can be obtained by simply computing this quantity for each feature, an operation of trivial complexity [16], and ordering the features by its (decreasing) value. This is the essence of FS by maximum marginal diversity (MMD), and the algorithm that was introduced in [16].

While the computational simplicity of MMD is quite appealing, its effectiveness for recognition problems will depend on the validity of the assertion, implicit in (4), that (on average) feature dependencies do not provide information about the class label. This is a difficult assertion to prove, since a precise characterization of how well (4) holds would require the accurate estimation of the joint densities of a large number of features. Such estimation seems infeasible, given the well known limitations of density estimation in high-dimensional spaces.

3. A family of ITFS algorithms

In this paper, we address this question through an alternative, indirect, strategy, based on the sequential relaxation of the assumption that feature dependencies are non-informative (with regards to the class label). For this, we start by grouping the features into a collection of disjoint subsets. The features within each subset are allowed to have arbitrary (i.e. informative) dependences, while the dependences between the subsets are constrained to be non-informative. By gradually increasing the cardinality of these subsets we move from the scenario where we have a large collection of individual features that all depend in a non-informative way (the scenario where MMD is optimal), to one where we have a single set of features that all depend in informative ways (the completely unconstrained scenario). We then extend (5) to account for each of these scenarios, and obtain the associated sequence of optimal FS algorithms.

This strategy is interesting in two ways. First, by applying these algorithms to real recognition tasks and measuring the error rates associated with the resulting sequence of feature spaces, we can identify the cardinality at which the assumption of non-informative dependences between feature subsets becomes realistic. Hopefully this cardinality will be small, enabling optimal FS with reduced computational complexity. In this regard, while the MMD assertion that all dependencies are non-informative may be too restrictive, one would expect this property to hold for at least some of the dependences. For example, the fact that the correlation between wavelet coefficients is significant for immediate neighbors (in both space and scale), rapidly decaying for coefficients at very different scales or orientations, hints that most pairs of coefficients are not jointly informative about the image class. Second, the resulting family of algorithms enables a continuous trade-off between computation and optimality. If the cardinality at which the inter-subset dependencies cease to be informative is N, algorithms that assume a (gradually) smaller cardinality will be (increasingly) sub-optimal but computationally more efficient.

3.1. l-decomposability

We start by introducing the concept of a $l-decomposable$ feature set.

Definition 1 *Let $\mathbf{X} = (X_1, \ldots, X_N)$ be a feature set of size N. The set $\mathbf{X}$ is l-decomposable, or decomposable at order l, if and only if there is a set of mutually exclusive feature subsets $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_{\lceil N/l \rceil}\}$ such that*

$$\mathbf{C}_i = \begin{cases} \{X_{(i-1)l+1}, \ldots, X_{il}\}, & \text{if } i < \lceil N/l \rceil, \\ \{X_{(i-1)l+1}, \ldots, X_N\}, & \text{if } i = \lceil N/l \rceil \end{cases}$$

and, for all $k \in \{2, \ldots, N\}$,

$$\sum_{i=1}^{\lceil k-1/l \rceil} \Big[I(X_k; \tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1, \ldots, \mathbf{C}_{i-1}) - I(X_k; \tilde{\mathbf{C}}_{i,k}) \Big] = \tag{6}$$
$$\sum_{i=1}^{\lceil k-1/l \rceil} \Big[I(X_k; \tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1, \ldots, \mathbf{C}_{i-1}, Y) - I(X_k; \tilde{\mathbf{C}}_{i,k}|Y) \Big].$$

where $\tilde{\mathbf{C}}_{i,k}$ is the subset of $\mathbf{C}_i$ containing the features of index smaller than k.

From the well known property that if A and B are jointly independent of C then $I(A,B|C) = I(A,B)$ it follows that the quantity $\frac{1}{\lceil k-1/l \rceil}\sum_{i=1}^{\lceil k-1/l \rceil}\left[I(X_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1}) - I(X_k;\tilde{\mathbf{C}}_{i,k})\right]$ is a measure of the average redundancy between the feature subsets $\mathbf{C}_i$. $\mathbf{X}$ is $l-decomposable$ if this average redundancy is non-informative with respect to the class label. Note that unlike (4), *l-decomposability does not impose any constraints on the dependencies between features that belong to the same subset*. The following proposition shows that when arbitrary dependencies of order l are allowed, the optimal infomax FS solution only requires density estimates on subspaces of dimension $l+1$.

Proposition 1 *Let $\mathbf{X}^* = (X_1^*,\ldots,X_N^*)$ be the optimal feature subset of size N, in the infomax sense. If $\mathbf{X}^*$ is l-decomposable into the set $\mathcal{C} = \{\mathbf{C}_1,\ldots,\mathbf{C}_{\lceil N/l \rceil}\}$ then*

$$I(\mathbf{X}^*;Y) = \sum_{k=1}^{N} I(\mathbf{X}_k^*;Y) - \sum_{k=2}^{N}\sum_{i=1}^{\lceil k-1/l \rceil}[I(X_k^*;\tilde{\mathbf{C}}_{i,k}) - I(X_k^*;\tilde{\mathbf{C}}_{i,k}|Y)] \quad (7)$$

where $\tilde{\mathbf{C}}_{i,k}$ is the subset of $\mathbf{C}_i$ containing the features of index smaller than k.

Proof: see the appendix.

3.2. Algorithms

The order l therefore ties the computational complexity of optimal FS to the subset cardinality required to achieve the optimal solution, i.e. the one which accounts for all the feature dependences that are informative with respect to the class label. This suggests the following family of algorithms that gradually trade optimality for speed (as l decreases).

Algorithm 1 *Given a set of n features $\mathbf{X} = (X_1,\ldots,X_n)$, the order l, the target number of features N, and denoting the marginal diversity of X_k, $I(\mathbf{X}_k,Y)$, by md_k.*

1. *set $\mathbf{X}^* = \mathbf{C}_1 = \{X_1^*\}$ where $X_1^* \in \mathbf{X}$ is the feature of largest marginal diversity, set $k=2$, and $i=1$.*
2. *foreach $X_r \notin \mathbf{X}^*$, compute $\delta_r = \sum_{p=1}^{\lceil k-1/l \rceil} I(X_r;\tilde{\mathbf{C}}_{p,k}) - I(X_r;\tilde{\mathbf{C}}_{p,k}|Y)$.*
3. *let $r^* = \arg\max_r md_r - \delta_r$. If $k-1$ is not a multiple of l make $\mathbf{C}_i = \mathbf{C}_i \cup \mathbf{X}_{r^*}$. Else, set $i = i+1$, and let $\mathbf{C}_i = \mathbf{X}_{r^*}$. In both cases make $\mathbf{X}^* = \cup_i \mathbf{C}_i$, $k = k+1$, and go to 2 if $k < N$.*
4. *return $\mathbf{X}^*$.*

The algorithms in this family belong to the class of forward search FS methods [21]. At each step the best feature, in the sense of (7), that has not yet been selected is identified and added to the selected set. If this contains a feature subset with less than l features the new feature is included in that subset. Otherwise, a new subset is created. The feature selection cost $md_r - \delta_r$ favors features that are discriminant (large md_r) but penalizes features that, when combined with those already in the selected set, are highly informative about the class label Y (large δ_r). The overall complexity is determined by the loop in step 4. If there are C classes, F feature vectors per class, and all densities are estimated with histograms of b bins along each axis, this involves, for the i^{th} subset $\mathbf{C}_i$, 1) estimating the joint histogram between X_r and the variables in $\mathbf{C}_i$, for all classes - complexity $O(FlC)$ - 2) using those histograms to compute $I(X_r;\mathbf{C}_i|Y)$ - complexity $O(b^lC)$ - and 3) their average to compute $I(X_r;\mathbf{C}_i)$ - complexity $O(b^lC)$. The entire loop has complexity $O[(F+b^l/l)kC]$ and, since it is repeated $n-k$ times, the complexity of step 4 is $O[(F+b^l/l)k(n-k)C]$. Summing from $k=1$ to N leads to an overall complexity of $O[n(F+b^l/l)N^2C]$. Since N is usually small it follows that the complexity is determined by the number of bins b and the order l.

4. Experimental results

In this section we report on a collection of experiments designed to evaluate various properties of ITFS. In particular, we considered four questions of practical significance: 1) is there a noticeable gain, in terms of probability of error, associated with modeling feature redundancies, 2) does that gain justify the computational cost inherent to higher dimensional density estimation, 3) what dependency order achieves the optimal trade-off between recognition accuracy and complexity, and 4) how does the recognition performance compare with standard variance-based methods, such as PCA, that address redundancy reduction but are not necessarily discriminant? The experiments were conducted on both the Brodatz (112 classes, 1008 images) and the Corel (15 classes, 1500 images) image databases, using a set up identical to that reported in some of our previous work (e.g. see [23]). In a nutshell, the database is divided into a training and test set (roughly 80 to 20%), the training set used for all the learning and the test set for evaluation. This consists of using the training images as a database, the test images as queries, evaluating the retrieval precision and recall (PR) for each query, and averaging over all queries. To measure the dependence of retrieval accuracy on the number of selected features, the average PR curve was summarized by its integral, the PR area (PRA). In

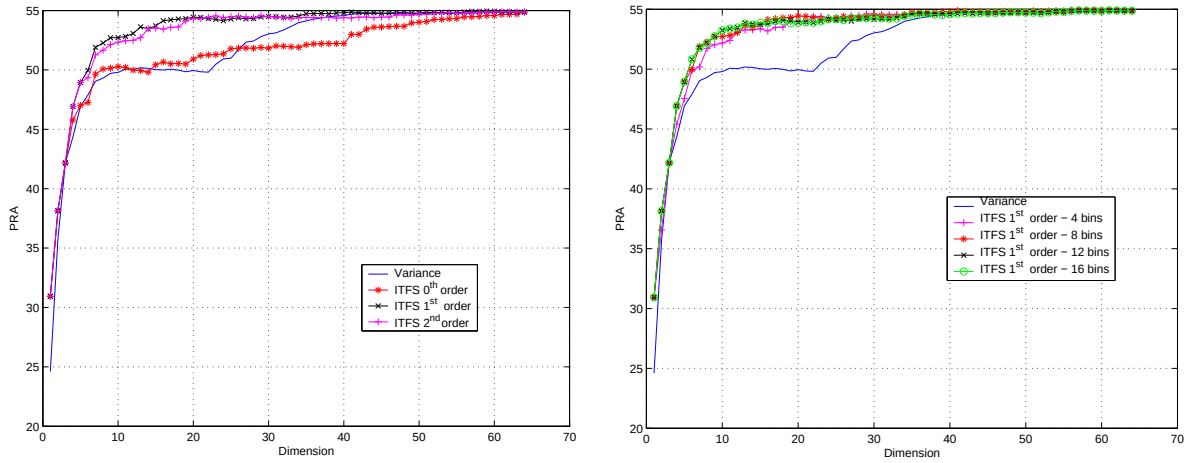

Figure 1: Left: PRA curves for the DCT feature set in Corel using ITFS algorithms of order 0 ('*'), 1 ('x'), and 2 ('+'), and energy compaction (solid line). Right: variation of the PRA curve with the number of histogram bins b.

all experiments feature vectors were extracted from random 8×8 image neighborhoods and all classification/retrieval results obtained with classifiers based on Gaussian mixtures. Various feature transformations were considered: the discrete cosine transform (DCT), a wavelet representation (WAV), principal component analysis (PCA), and independent component analysis (ICA). Figure 1 (left) presents the PRA curves obtained on Corel with the DCT features (similar results obtained with the remaining transformations as well as on Brodatz are omitted for brevity) for $l \in \{0, 1, 2\}$ ($l = 0$ corresponding to the maximization of the marginal diversity as given by (5)). These curves are compared to that obtained with the same features and the energy compaction criteria. It is visible that 1^{st}-order decomposability is sufficient to guarantee very substantial improvements over energy compaction. For example, in the DCT case, it takes energy compaction 40 features to reach the accuracy that ITFS achieves with only 10! In fact, even the simple maximization of marginal diversity compares well to energy compaction, achieving higher accuracy when the number of selected features is small, the situation of greatest practical interest. Finally, the assumption of 2^{nd} order decomposability does not lead to any increase in PRA over that of 1^{st} order. This suggests that all the feature dependencies that matter for recognition are those of first order, a significant result given the exponential dependence of the computational complexity on l.

To further analyze the dependence of recognition accuracy on computational complexity, we repeated all experiments for different numbers of histogram bins b. Figure 1 (right) presents the resulting PRA curves when $l = 1$. Clearly, no noticeable changes happen above $b = 8$ bins, and the PRA is very close to the best even when $b = 4$. This indicates that the performance of the ITFS algorithms is quite insensitive to the number of histogram bins and coarse histograms are sufficient to guarantee good retrieval results. When combined with the left plot this result shows that ITFS enables significant improvements over energy compaction at the expense of a small increase in complexity. Overall, the best trade-off between accuracy and complexity is achieved by the ITFS algorithm with $l = 1$. Visual inspection of the recognition results shows a significant improvement for queries from classes that have visual attributes in common with other classes in the database. Since features of large variance are not necessarily discriminant they lead to confusion between such classes, as is illustrated by the left column of Figure 2. For example, in the query displayed in the top row, images from the flower class are confused with images from the horses class, because the backgrounds are similar in the two classes. The picture on the right column shows that, when the discriminant ITFS features are used, the retrieval precision increases significantly. Similar results are presented in the subsequent rows for queries involving various other overlapping class pairs, namely fireworks vs stained glass, coasts vs scuba diving scenes, ski scenes vs glaciers and mountains, and monuments vs oil paintings. In general, we have observed that in queries for images from classes that have significant overlap with other classes in the database (i.e. the most difficult queries) ITFS leads to significantly higher retrieval accuracy than the variance-based criteria.

References

[1] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image Classification for Content-Based Indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, January 2001.

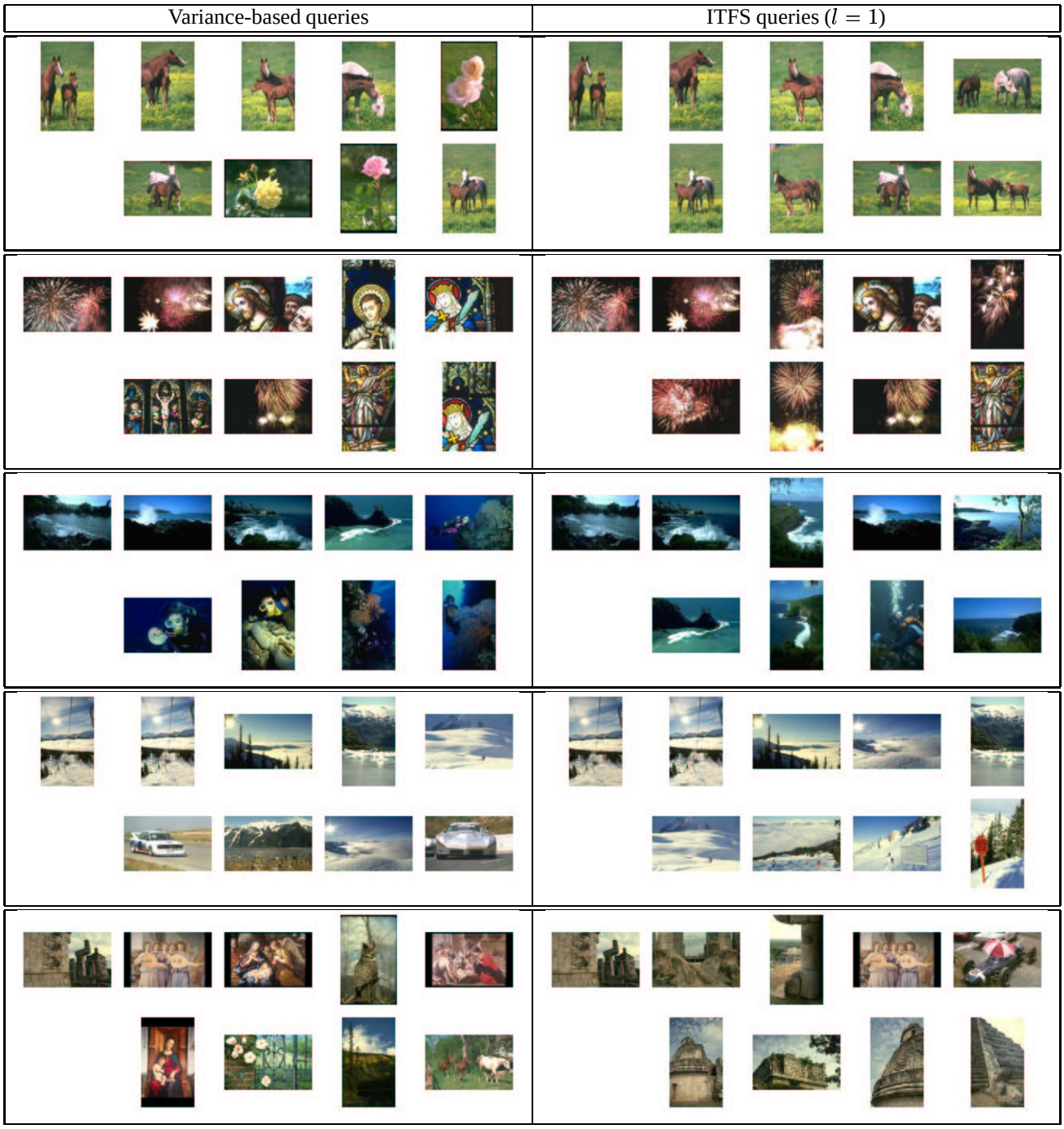

Figure 2: From top to bottom: five queries from Corel. Left: 8 best matches when FS is based on variance. Right: 8 best matches under ITFS. In both cases the query image is shown on the top-left corner.

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

[3] B. Manjunath and W. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 18(8):837–842, August 1996.

[4] M. Do and M. Vetterli. Wavelet-based Testure Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. *IEEE Trans. on Image Processing*, Vol. 11(2):146–158, February 2002.

[5] J. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Columbia University, 1997.

[6] Yong Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998.

[7] M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, Vol. 7(1):11–32, 1991.

[8] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Spatial Color Indexing and Applications. *Int. Journal of Computer Vision*, 35(3):245–268, December 1999.

[9] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. In *International Conference on Computer Vision, Korfu, Greece*, pages 1165–1173, 1999.

[10] I. Cox, M. Miller, S. Omohundro, and P. Yianilos. PicHunter: Bayesian Relevance Feedback for Image Retrieval. In *Int. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

[11] A. Jain and A. Vailaya. Image Retrieval using Color and Shape. *Pattern Recognition Journal*, 29:1233–1244, August 1996.

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

[13] J. De Bonet and P. Viola. Structure Driven Image Database Retrieval. In *Neural Information Processing Systems, Denver, Colorado*, volume 10, 1997.

[14] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Color- and Texture-Based Image Segmentation Using EM and Its Application to Image Querying and Classification. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, Vol. 8(24):1026–1038, August 2002.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[21] 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.

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

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

A. Proof of Proposition 1

Proof: By recursive application of the chain rule of mutual information

$$\begin{aligned} I(X_k^*; \mathbf{X}_{1,k-1}^* | Y) &= I(X_k^*; \mathbf{C}_1, \ldots, \tilde{\mathbf{C}}_{\lceil k-1/l \rceil, k} | Y) \\ &= I(X_k^*; \tilde{\mathbf{C}}_{\lceil k-1/l \rceil, k} | \mathbf{C}_1, \ldots, \mathbf{C}_{\lceil k-1/l \rceil - 1}, Y) + \\ &\quad I(X_k^*; \mathbf{C}_1, \ldots, \mathbf{C}_{\lceil k-1/l \rceil - 1} | Y) \\ &= \sum_{i=1}^{\lceil k-1/l \rceil} I(X_k^*; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1, \ldots, \mathbf{C}_{i-1}, Y). \end{aligned}$$

and, similarly,

$$I(X_k^*;\mathbf{X}_{1,k-1}^*) = \sum_{i=1}^{\lceil k-1/l \rceil} I(X_k^*;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1})$$

It follows from (3) that

$$I(\mathbf{X}^*;Y) = \sum_{k=1}^{N} I(\mathbf{X}_k^*;Y) + \tag{8}$$

$$+ \sum_{k=2}^{N}\left[\sum_{i=1}^{\lceil k-1/l \rceil} I(X_k^*;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1},Y)\right. \tag{9}$$

$$\left. - \sum_{i=1}^{\lceil k-1/l \rceil} I(X_k^*;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1})\right]$$

and, rewriting (6) as

$$\sum_{i=1}^{\lceil k-1/l \rceil} \Big[I(X_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1}) - I(X_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1,\ldots,\mathbf{C}_{i-1},Y)\Big] =$$

$$\sum_{i=1}^{\lceil k-1/l \rceil} \Big[I(X_k;\tilde{\mathbf{C}}_{i,k}) - I(X_k;\tilde{\mathbf{C}}_{i,k}|Y)\Big],$$

(7) follows from the fact that $\mathbf{X}^*$ is l-decomposable into $\mathcal{C} = \{\mathbf{C}_1,\ldots,\mathbf{C}_{\lceil N/l \rceil}\}$.