Natural Image Statistics and Low-Complexity Feature Selection

Manuela Vasconcelos and Nuno Vasconcelos, *Senior Member*, *IEEE*

Abstract—Low-complexity feature selection is analyzed in the context of visual recognition. It is hypothesized that high-order dependences of bandpass features contain little information for discrimination of natural images. This hypothesis is characterized formally by the introduction of the concepts of conjunctive interference and decomposability order of a feature set. Necessary and sufficient conditions for the feasibility of low-complexity feature selection are then derived in terms of these concepts. It is shown that the intrinsic complexity of feature selection is determined by the decomposability order of the feature set and not its dimension. Feature selection algorithms are then derived for all levels of complexity and are shown to be approximated by existing information-theoretic methods, which they consistently outperform. The new algorithms are also used to objectively test the hypothesis of low decomposability order through comparison of classification performance. It is shown that, for image classification, the gain of modeling feature dependencies has strongly diminishing returns: best results are obtained under the assumption of decomposability order 1. This suggests a generic law for bandpass features extracted from natural images: that the effect, on the dependence of any two features, of observing any other feature is constant across image classes.

Index Terms—Feature extraction and construction, low complexity, natural image statistics, information theory, feature discrimination versus dependence, image databases, object recognition, texture, perceptual reasoning.

1 Introduction

Natural image statistics have been a subject of substantial recent research in computer and biological vision [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. For computer vision, good models of image statistics enable algorithms tuned to the scenes that matter the most. Tuning to natural statistics can be accomplished through priors that favor solutions consistent with them [15], [16], [17], [18] or through optimal solutions derived from probability models that enforce this consistency [19], [20], [21], [22], [23]. The idea of optimal tuning to natural statistics also has a long history in biological vision [5], [24], [25], [26], where this tuning is frequently used to justify neural computations. In fact, various recent advances in computational modeling of biological vision have followed from connections between neural function and properties of natural stimulus statistics [8], such as sparseness [12], [27], independence [13], [14], compliance with certain probability models [28], or optimal statistical estimation [29], [30].

Although natural images are quite diverse, their convolution with banks of bandpass functions gives rise to frequency *coefficients* with remarkably stable statistical properties [1], [2], [3], [4], [6], [7], [8], [10]. This is illustrated in Fig. 1a, which presents three images, the histograms of one coefficient of their wavelet decomposition, and the histograms of that coefficient conditioned on its parent. The different visual appearance of the images affects the scale (variance) of the marginal distribution *but not its shape or that of the conditional distribution*, which is a bow-tie for all classes. This canonical pattern is simply rescaled to match the marginal statistics of each class. These types of properties have been exploited in various image processing domains, including compression [1], [2], [6], [19], denoising [15], [16], [18], [22], retrieval [21], saliency [31], extraction of intrinsic images [20], separation of reflections [32], and inpainting [17], [18]. In fact, the study of image statistics has a complementary relationship with the development of vision algorithms. Typically, an hypothesis is advanced for the statistics, an algorithm is derived under that hypothesis, and applied to natural images. If the algorithm performs well, the hypothesis is *validated*.

This *indirect validation paradigm* is useful in two ways. First, it avoids the estimation of complex statistical quantities. For example, hypotheses on high-order statistics are difficult to verify experimentally due to the well-known difficulties of estimating such statistics [33]. Instead, it is usually easier to 1) derive an algorithm that is *optimal* if the hypothesis holds and 2) apply it to a specific vision problem such as object recognition [34], where performance can be *easily quantified*. If the algorithm performs poorly, there is reason to question the hypothesis; otherwise, there is concrete evidence in its support. The second advantage of indirect validation is that it produces new vision algorithms which, *under the hypothesis*, are *optimally* tuned to the image statistics. If the hypothesis holds, these algorithms can outperform the state of the art.

In this work, we adopt the indirect validation paradigm to study the *discriminant power of the statistical dependencies of frequency coefficients extracted from natural images*. While simple inspection of the histograms in Fig. 1a shows that these dependences exist, their constancy across image classes suggests the hypothesis that *high-order dependences*

- *The authors are with the Statistical Visual Computing Laboratory, University of California, San Diego, 9500 Gilman Drive, MC 0407, La Jolla, CA 92093. E-mail: maspcv@gmail.com, nuno@ece.ucsd.edu.*

Manuscript received 16 Nov. 2006; revised 16 Oct. 2007; accepted 1 Feb. 2008; published online 24 Mar. 2008.
Recommended for acceptance by B.S. Manjunath.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0813-1106.
Digital Object Identifier no. 10.1109/TPAMI.2008.77.

0162-8828/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society

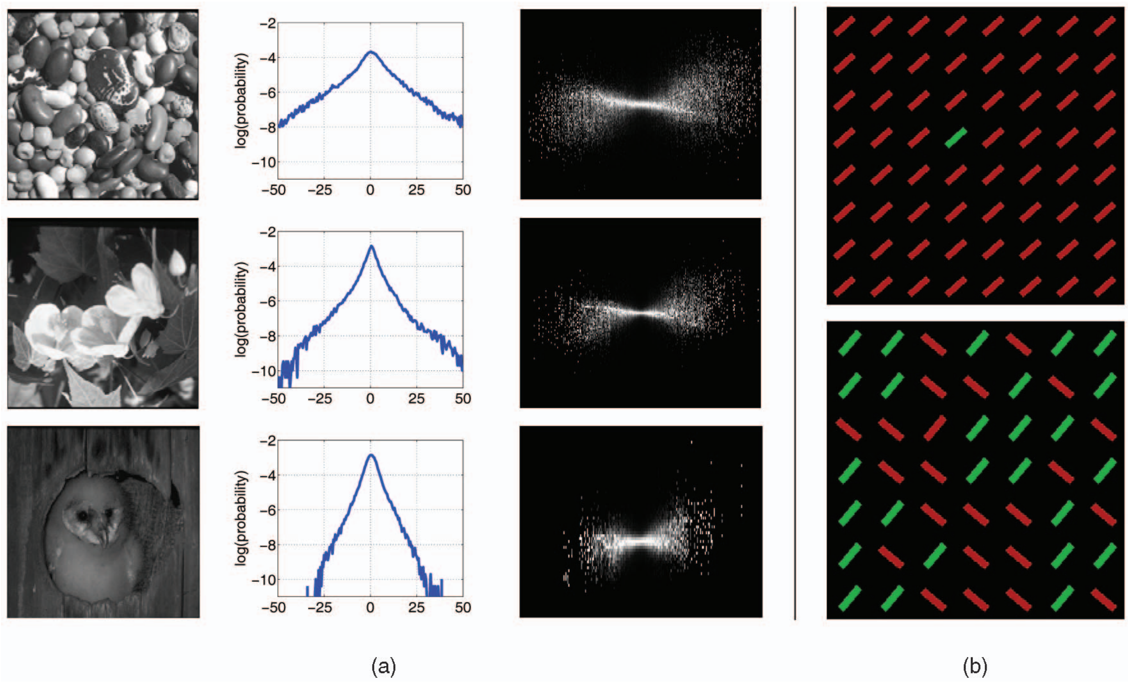

Fig. 1. (a) Constancy of natural image statistics. (a.1) Three images. (a.2) Each plot presents the histogram of the same coefficient from a wavelet decomposition of the image on the left. (a.3) Conditional histogram of the coefficient conditioned on the value of the colocated coefficient of an immediately coarser scale (its parent). (b) Biological vision frequently disregards feature dependences. (b.1) A Stimulus that differs from it surrounds by a single feature (color) is salient. (b.2) Differences in feature conjunctions (color and orientation) are not.

contain little information for image discrimination. This hypothesis is supported by what is known about biological vision, where it has long been argued that the early visual system dismisses feature dependences in the solution of discriminant tasks such as visual search [35], [36]. This is illustrated in Fig. 1b, which presents a classical example of the inability of preattentive vision to process feature conjunctions. When, as shown in Fig. 1b.1, an object (colored bar) differs from a background of distractors (other colored bars) in terms of a single feature (color), it can be easily discriminated (it *pops out*). However if, as shown in Fig. 1b.2, the object differs from the distractors by a conjunction of two features (color and orientation, the bar on the third row and third column), there is no percept of *popout*. Current explanations attribute this phenomena to independent feature processing [35], [36], [37], [38], [39], [40].

For computer vision, where models of feature dependences require the estimation of high-dimensional densities, such dependences are a dominant source of complexity. A formal characterization of their role in image discrimination is therefore a prerequisite for *optimal image classification with reduced complexity*. Since optimal classification requires discriminant features, we study dependences in the context of feature selection. In the spirit of indirect validation, we 1) develop optimal feature selection algorithms under the hypothesis that high-order dependences are uninformative for discrimination and 2) evaluate their image classification performance.

The contributions of this effort are in three areas. The first is a rigorous characterization of the role of image statistics in optimal feature selection with low complexity. We equate complexity with the dimensionality of the probability densities to be estimated, and adopt an information-theoretic definition of optimality widely used in the literature [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65]. We then derive, for each level of complexity, *the necessary and sufficient condition (on the statistics) for optimal feature selection with that complexity*. This condition depends exclusively on a quantity denoted as the *conjunctive interference* within the set of features $\mathbf{X}$, which roughly measures how, on average, the dependence between two disjoint feature subsets $\mathbf{A}, \mathbf{B} \subset \mathbf{X}$ is affected by the remaining features in $\mathbf{X}$. It is shown (see Theorem 1) that if this conjunctive interference is constant across classes, the complexity of the optimal solution is determined by the dimension of the subsets $\mathbf{A}$, $\mathbf{B}$ rather than that of $\mathbf{X}$. Hence, *the smaller the set size for which conjunctive interference is nondiscriminant, the smaller the intrinsic complexity of feature selection*.

The second contribution, which follows from the theoretical analysis, is a new family of feature selection algorithms. These algorithms optimize simplified costs at all levels of complexity and are (locally) optimal when conjunctive interference is nondiscriminant at their complexity level. This family generalizes a number of low-complexity information-theoretic methods [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64] previously shown to outperform many state-of-the-art feature selection techniques [48], [58]. The impressive empirical performance of the previous methods is explained by the fact that they *approximate* the algorithms now derived. Nevertheless, there is a gain in replacing the approximations with the optimal algorithms: experiments on various data sets show that the latter *consistently outperform the previous methods*, sometimes by a significant margin.

The final contribution, in the spirit of indirect validation, is the use of the feature selection algorithms for indirectly characterizing the image statistics. Given that the different algorithms are optimal only when conjunctive interference is nondiscriminant at their complexity level, a comparison of feature selection performance identifies the complexity at which conjunctive interference ceases to affect image discrimination. Algorithms with less than this complexity are suboptimal, and performance levels off once it is reached. We present evidence for the hypothesis that this "leveling off" effect occurs at very low complexity levels. While simply modeling marginal densities is, in general, not enough to guarantee optimal feature selection, *there appears to be a little gain in estimating more than the densities of pairs of coefficients*.

This paper is organized as follows: Section 2 reviews information-theoretic feature selection. Section 3 introduces a basic decomposition of the information-theoretic cost and shows that independent feature selection can be optimal, even for highly dependent feature sets. The decomposition is refined in Section 4, which formally defines conjunctive interference, and introduces a measure of the intrinsic complexity of a feature set (*decomposability order*). Section 5 introduces the new family of (locally) optimal algorithms and discusses connections to prior methods. Finally, the experimental protocol for indirect validation of the decomposability hypothesis is introduced in Section 6, and experimental results are discussed in Section 7. A very preliminary version of the work, focusing mostly on the theoretical connections between conjunctive interference and low complexity feature selection, has appeared in [64].

2 Infomax Feature Selection

We start by introducing the information-theoretic optimality criterion adopted in this work and reviewing its previous uses in the feature selection literature.

2.1 Definitions

A classifier $g : \mathcal{X} \rightarrow \mathcal{L} = \{1, \dots, M\}$ maps a feature vector $\mathbf{x} = (x_1, \dots, x_N)^T \in \mathcal{X} \subset \mathbb{R}^N$ into a class label $i \in \mathcal{L}$. Feature vectors result from a transformation $T : \mathcal{Z} \rightarrow \mathcal{X}$ of observation vectors $\mathbf{z} = (z_1, \dots, z_D)$ in measurement space $\mathcal{Z} \subset \mathbb{R}^D$. Observations are samples from random process $\mathbf{Z}$ of probability distribution $P_{\mathbf{Z}}(\mathbf{z})$ on $\mathcal{Z}$, feature vectors samples from process $\mathbf{X}$ of distribution $P_{\mathbf{X}}(\mathbf{x})$ on $\mathcal{X}$, and label samples from random variable Y of distribution $P_Y(i)$ in $\mathcal{L}$. Given class i, observations have class-conditional density $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$ and class-posterior probabilities determined by the Bayes rule $P_{Y|\mathbf{Z}}(i|\mathbf{z}) = P_{\mathbf{Z}|Y}(\mathbf{z}|i)P_Y(i)/P_{\mathbf{Z}}(\mathbf{z})$. The classification problem is uniquely defined by $\mathcal{C} = \{\mathcal{Z}, P_{\mathbf{Z}|Y}(\mathbf{z}|i), P_Y(i), i \in \mathcal{L}\}$. T induces class-conditional densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ in $\mathcal{X}$ and defines a new classification problem $\mathcal{C}_{\mathcal{X}} = \{\mathcal{X}, P_{\mathbf{X}|Y}(\mathbf{x}|i), P_Y(i), i \in \mathcal{L}\}$. We define as optimal the spaces of maximum mutual information (MI) between features and class labels.

Definition 1. *Given a classification problem $\mathcal{C}$ and a set $\mathcal{S}$ of range spaces for the feature transforms under consideration, the infomax space is*

$$\mathcal{X}^* = \arg\max_{\mathcal{X} \in \mathcal{S}} I(Y; \mathbf{X}), \tag{1}$$

where

$$I(\mathbf{X}; Y) = \sum_i \int_{\mathcal{X}} p_{\mathbf{X},Y}(\mathbf{x}, i) \log \frac{p_{\mathbf{X},Y}(\mathbf{x}, i)}{p_{\mathbf{X}}(\mathbf{x}) p_Y(i)} d\mathbf{x} \tag{2}$$

is the MI between $\mathbf{X}$ and Y.

Infomax is closely related to the minimization of Bayes classification error and has a number of relevant properties for low-complexity feature selection, some of which are reviewed in Appendix A. In what follows, $\mathbf{z}$ is a vector of image pixels, and $\mathbf{x}$ is the result of a bandpass transformation (e.g., a wavelet, Gabor, or windowed Fourier transform), followed by the selection of N coefficients.

2.2 Previous Infomax Approaches to Feature Selection

Information-theoretic feature selection has been used for text categorization [41], [42], [43], [44], creation of semantic ontologies [45], analysis of genomic microarrays [46], [47], classification of electroencephalograms (EEGs) [49], [50] and sonar pulses [53], [54], medical diagnosis [51], audio-visual speech recognition [56], and visualization [57]. In computer vision, it has been used for face detection [58], object recognition [59], [61], and image retrieval [62], [63], [64]. These approaches can be grouped into four classes. Algorithms in the first class approximate (2) with

$$M(\mathbf{X}; Y) = \sum_{k=1}^{D} I(X_k; Y), \tag{3}$$

where $I(X_k; Y)$ is the MI between feature X_k and class label Y. $M(\mathbf{X}; Y)$ is a measure of the discriminant information conveyed by individual features. It is denoted as the *marginal MI* (MMI), and its maximization is denoted as *marginal infomax*. It is popular in text categorization [41], [42], [43] mostly due to its computational simplicity. It has, nevertheless, been shown to sometimes outperform methods that account for feature dependences [45], [51], [56].

Algorithms in the second class combine a heuristic extension of marginal infomax, originally proposed in [53], and the classical greedy strategy of sequential forward feature selection [66], where one feature is selected at a time. Denoting by $\mathbf{X}^* = \{X_1^*, \dots, X_k^*\}$ the set of previously selected features and denoting by X a candidate feature, the selected feature is

$$X_{k+1}^* = \arg\max_X \{I(X; Y) - f(X, \mathbf{X}^*)\}, \tag{4}$$

where $f(\cdot)$ is a dependence measure, ranging from a hard rejection of dependent features [53] to continuous penalties. The most popular is [47], [48], [51], [52], [54], [55]

$$f(X, \mathbf{X}^*) = \xi \sum_{i=1}^{k} I(X; X_i^*), \tag{5}$$

where ξ controls the strength of the dependence penalty. Various information-theoretic costs are either special cases of this [47], [48] or extensions that automatically determine ξ [55].

Algorithms in the third class optimize costs closer to (1), once again through sequential forward search. One proposal is to select the feature X, which maximizes $I(X, X_i^*; Y)$, $i \in \{1, \dots, k\}$ [57]. This is a low-complexity approximation to $I(\mathbf{X}; Y)$, which only considers pairs of features. Because it does not rely on a modular decomposition of the MI, it is somewhat inefficient. An alternative, proposed in [58] and [60], addresses this problem by relying on

$$\begin{aligned} X_{k+1}^* &= \arg\max_X \min_i I(Y; X|X_i^*) \\ &= \arg\max_X \min_i \big[I(X, X_i^*; Y) - I(X_i^*; Y)\big], \end{aligned} \tag{6}$$

where we have used (31). This is equivalent (see (34)) to

$$X_{k+1}^* = \arg\max_X \{I(X; Y) + \min_i \big[I(X; X_i^*|Y) - I(X; X_i^*)\big]\}. \tag{7}$$

We will show that (4) and (7) are simplifications of (1), which disregard important components for image discrimination. Nevertheless, extensive empirical studies have shown that they can beat state-of-the-art methods [48], [58] such as boosting [67], [68] and decision trees [69].

The final class has a single member, i.e., the algorithm in [65]. Unlike the other classes, it sequentially eliminates features from $\mathbf{X}$. This elimination is based on the concept of a Markov blanket [70]: if there is a set of features $\mathbf{M}$ (called a Markov blanket) such that X is conditionally independent of $(\mathbf{X} \cup Y) - \mathbf{M} - \{X\}$, given $\mathbf{M}$, the feature X can be removed from $\mathbf{X}$ without any loss of information about Y. While theoretically sound, this method has a number of practical shortcomings that are acknowledged by its authors: the Markov blanket condition is much stronger than what is really needed (conditional independence of X from Y, given $\mathbf{M}$), there may not be a full Markov blanket for a feature, and when there is one, it can be difficult to find. To overcome these problems, Koller and Sahami [65] use various heuristics that only involve feature pairs. The assumptions, with respect to the feature statistics, underlying these heuristics are not clear.

3 Optimality of Marginal Infomax

To gain some intuition on the feasibility of low-complexity feature selection, we start by investigating the conditions under which marginal infomax is identical to (1).

3.1 Features versus Conjunctions

For this, we note that the MI can be decoupled into contributions from individual features and feature conjunctions.

Lemma 1. *Let $\mathbf{X} = (X_1, \dots, X_D)$ be any feature set and let $\mathbf{X}_{1,k} = (X_1, \dots, X_k)$. Then*

$$I(\mathbf{X}; Y) = M(\mathbf{X}; Y) + C(\mathbf{X}; Y), \tag{8}$$

where $M(\mathbf{X}; Y)$ is the MMI of (3), and

$$C(\mathbf{X}; Y) = \sum_{k=2}^{D} [I(X_k; \mathbf{X}_{1,k-1}|Y) - I(X_k; \mathbf{X}_{1,k-1})]. \tag{9}$$

Proof. See Appendix B. □

The terms $I(X_k; \mathbf{X}_{1,k-1}|Y) - I(X_k; \mathbf{X}_{1,k-1})$ measure how the MI between features is affected by knowledge of the class label. They quantify the discriminant information due to feature dependences. $C(\mathbf{X}; Y)$ is referred to as the *conjunctive component* of the MI (CCMI). A consequence of Lemma 1 is that if $C(\mathbf{X}, Y) = 0, \forall \mathcal{X} \in \mathcal{S}$, then (1) reduces to the marginal infomax criterion

$$\mathcal{X}^* = \arg\max_{\mathcal{X} \in \mathcal{S}} \sum_k I(X_k; Y). \tag{10}$$

Due to the nonnegativity of the MI, (10) has a simple solution: order the X_k by decreasing $I(X_k; Y)$ and select the largest N. While (1) involves combinatorial search and high-dimensional density estimation, (10) only requires a linear search based on marginal density estimates. Hence, a null CCMI is a sufficient condition for low-complexity feature selection.

3.2 The Role of Natural Image Statistics

To obtain some intuition on how the CCMI is affected by the dependency structure of $\mathbf{X}$, we consider the classification of two Gaussian features $\mathbf{X} = (X_1, X_2)$ with

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \frac{1}{\sqrt{4\pi^2|\mathbf{\Sigma}_i|}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}_i^{-1} \mathbf{x}}, \; i \in \{1, 2\},$$

$$\mathbf{\Sigma}_i = \begin{bmatrix} \epsilon_i & \gamma_i \\ \gamma_i & \eta_i \end{bmatrix}, \quad \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2.$$

Gaussianity reduces all class-conditional dependences to two parameters, namely, the correlation coefficients $\rho_i = \gamma_i / \sqrt{\epsilon_i \eta_i}$. It is relatively straightforward to measure the relative strength

$$R(\mathbf{X}; Y) = \frac{C(\mathbf{X}; Y)}{M(\mathbf{X}; Y)} \tag{11}$$

of the MI components as a function of these parameters. If the variances ϵ_i and η_i are held constant, fixing the marginal distributions, then $R(\mathbf{X}; Y)$ is proportional to $C(\mathbf{X}; Y)$, allowing for the study of how the latter depends on the ρ_i. By repeating the experiment with different ϵ_i and η_i, it is also possible to infer how this dependence is affected by the MMI $M(\mathbf{X}; Y)$. The graph of $R(\mathbf{X}; Y)$ versus ρ_i for fixed MMIs is the *CCMI surface* associated with the latter. While natural image statistics are not Gaussian, this procedure provides intuition on how the MI is affected by feature dependences. We consider two common scenarios for pairs of bandpass coefficients:

- **S1.** Two features are active/inactive for the same images (e.g., a wavelet coefficient and its parent). X_1 and X_2 have equal variance ($\epsilon_i = \eta_i = \nu_i$) and are inactive for one class ($\nu_2 = 1$) but are active for the other ($\nu_1 > 1$). The CCMI surface is measured for various activity levels (by controlling ν_1).

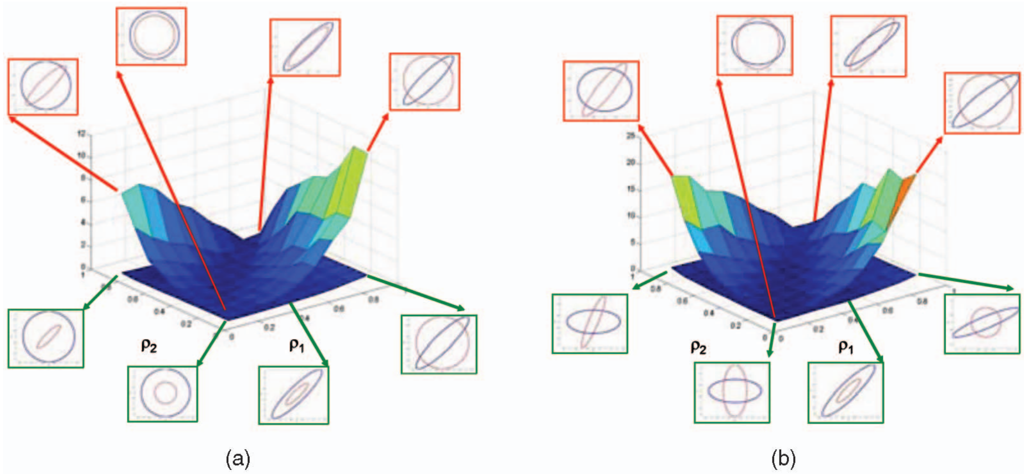

Fig. 2. $R(\mathbf{X}, Y)$ as a function of the class-conditional correlations ρ_i for a binary Gaussian problem. The inserts show the one standard deviation contour of the two Gaussian classes for various values of (ρ_1, ρ_2). The plots report to (a) scenario **S1** and (b) scenario **S2.** In both cases, different surfaces report to different values of ν, the variable that controls the marginal discrimination. All MIs were evaluated by replacing expectations with sample means, obtained from a sample of 10,000 points per class.

- **S2.** Each feature is active for one class but not for the other; for example, X_1 (X_2) is horizontally (vertically) tuned, and class 1 (2) is predominantly composed of horizontal (vertical) lines. The variances are $\epsilon_1 = \eta_2 = \nu$ and $\epsilon_2 = \eta_1 = 1$. The CCMI surface is measured for various ν.

Fig. 2 presents the corresponding CCMI surfaces, suggesting three main conclusions. First, *the CCMI can be close to zero, even when the features are very strongly dependent*. Note that all surfaces are approximately zero along the line $\rho_1 = \rho_2 = \rho$, independent of either ρ (dependence strength) or the MMI. Second, *the importance of the CCMI in (8) increases with the diversity of the dependence across classes*, i.e., with $|\rho_1 - \rho_2|$. Third, *this increase is inversely proportional to the MMI*. While, for small MMIs, a significant difference between the ρ_i makes $R(\mathbf{X}, Y)$ large, this is not the case for large MMIs. Overall, (8) (and Fig. 2) shows that 1) the relevance of feature dependences to the solution of (1) increases with their interclass variability but 2) this variability only boosts the importance of features that are not discriminant per se.

In summary, $C(\mathbf{X}, Y) = 0$ is a sufficient condition for optimal feature selection with low complexity. It *does not require feature independence but simply that the discriminant power of feature dependences is small*. As shown in Fig. 1a, this hypothesis is not unreasonable for natural images. We will evaluate it in Section 7. For now, we consider a series of extensions that bridge the gap between (1) and (10).

4 Decompositions of the Conjunctive Component

If feature conjunctions are discriminant, it is unlikely that this will hold for *all* conjunctions. For example, wavelet coefficients are dependent on their immediate neighbors (in space, scale, or orientation), but the dependence decays quickly [71]. Hence, $C(\mathbf{X}, Y)$ *should not require modeling dependences between all coefficients*. We next derive conditions for the optimality of infomax costs that only account for dependences within low-dimensional feature subsets.

4.1 Decompositions of the MI

We start by considering the decomposition of $I(\mathbf{X}, Y)$ for a given feature set $\mathbf{X}$. We group the D features into a collection of disjoint subsets of cardinality l:

$$\mathcal{C}_l = \{\mathbf{C}_1, \dots, \mathbf{C}_{\lceil D/l \rceil}\}, \tag{12}$$

where[1]

$$\mathbf{C}_i = \begin{cases} \{X_{(i-1)l+1}, \dots, X_{il}\}, & \text{if } i < \lceil D/l \rceil, \\ \{X_{(i-1)l+1}, \dots, X_D\}, & \text{if } i = \lceil D/l \rceil, \end{cases} \tag{13}$$

and $\lceil x \rceil = \inf\{m \in \mathbb{Z} | x \leq m\}$, and we derive the conditions under which the CCMI is totally determined by the dependencies within each $\mathbf{C}_i$. This is based on the following decomposition.

Lemma 2. *Consider the decomposition of* $\mathbf{X}$ *into a subset collection* $\mathcal{C}_l$*, as in (12). Then*

$$C(\mathbf{X}, Y) = \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}, Y) - I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}) \right], \tag{14}$$

where $\mathbf{C}_i$ *are as in (13),* $\tilde{\mathbf{C}}_{i,k}$ *is the subset of features in* $\mathbf{C}_i$ *whose index is smaller than* k*, and* $\mathbf{C}_1^{i-1} = (\mathbf{C}_1, \dots, \mathbf{C}_{i-1})$.

Proof. See Appendix C. □

This decomposition offers an explanation for why, in the absence of statistical regularities, low complexity feature selection is impossible [72]. Note that although $\mathbf{C}_1^{i-1}$ shares no elements with $\{X_k\}$ or $\tilde{\mathbf{C}}_{i,k}$, the state of the features of the former affects the dependences between those in the latter. Hence, *the discriminant information due to the dependences between* X_k *and* $\tilde{\mathbf{C}}_{i,k}$ *depends on the state of* $\mathbf{C}_1^{i-1}$ *and is impossible to compute with low complexity*. We refer to these indirect dependence relationships, i.e., that the state of a subset of features interferes with the dependence between two other

1. What follows could be extended to subsets $\mathbf{C}_i$ of different cardinality, but this would complicate the notation and is omitted.

nonoverlapping subsets, as *second-order components of dependence*. This is opposed to direct dependences between subsets, which are referred to as *first-order components* or dependences within subsets, which we denote as *zeroth order*. The *conjunctive interference* within a feature set is the overall difference between the first- and second-order dependences of its subsets.

Definiton 2. *Consider the decomposition of* $\mathbf{X}$ *into a subset collection* $\mathcal{C}_l$*, as in Lemma 2. The conjunctive interference within* $\mathbf{X}$ *with respect to* $\mathcal{C}_l$ *is*

$$CI(\mathbf{X};\mathcal{C}_l) = \sum_{k=2}^{D}\sum_{i=1}^{\lceil k-1/l\rceil}\left[I(X_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1^{i-1}) - I(X_k;\tilde{\mathbf{C}}_{i,k})\right]. \quad (15)$$

Conjunctive interference is a *differential measure of dependence*. It measures how, across the feature set, the dependence between two sets of features (e.g., $(X_k, \tilde{\mathbf{C}}_{i,k})$) *changes* with the observation of a third nonoverlapping set ($\mathbf{C}_1^{i-1}$). Since, if $(\mathbf{A},\mathbf{B})$ is independent of $\mathbf{C}$, $I(\mathbf{A};\mathbf{B}|\mathbf{C}) = I(\mathbf{A};\mathbf{B})$, it follows that conjunctive interference within $\mathbf{X}$ (with respect to decomposition $\mathcal{C}_l$) is null when $(X_k, \tilde{\mathbf{C}}_{i,k})$ is independent of $\mathbf{C}_1^{i-1}$ for all valid i and k. We next show that this is not a necessary condition for low-complexity evaluation of the MI. It suffices that the conjunctive interference does not depend on the class.

Theorem 1. *Consider the decomposition of* $\mathbf{X}$ *into* $\mathcal{C}_l$*, as in (12). Then,*

$$I(\mathbf{X};Y) = M(\mathbf{X};Y) + C_{\mathcal{C}_l}(\mathbf{X};Y), \quad (16)$$

with $M(\mathbf{X},Y)$ *as in (3), and*

$$C_{\mathcal{C}_l}(\mathbf{X};Y) = \sum_{k=2}^{D}\sum_{i=1}^{\lceil k-1/l\rceil}[I(X_k;\tilde{\mathbf{C}}_{i,k}|Y) - I(X_k;\tilde{\mathbf{C}}_{i,k})], \quad (17)$$

if and only if

$$CI(\mathbf{X};\mathcal{C}_l) = \sum_{k=2}^{D}\sum_{i=1}^{\lceil k-1/l\rceil}\left[I(X_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1^{i-1},Y) - I(X_k;\tilde{\mathbf{C}}_{i,k}|Y)\right]. \quad (18)$$

Proof. See Appendix D. □

When (18) holds, (16) is equivalent to (8), with $C_{\mathcal{C}_l}(\mathbf{X};Y)$ playing the role of $C(\mathbf{X};Y)$. In particular, (16) replaces each of the terms

$$I(X_k;\mathbf{X}_{1,k-1}|Y) - I(X_k;\mathbf{X}_{1,k-1}) \quad (19)$$

of (9) by a sum, over i, of terms of the form

$$I(X_k;\tilde{\mathbf{C}}_{i,k}|Y) - I(X_k;\tilde{\mathbf{C}}_{i,k}). \quad (20)$$

While (19) quantifies the discriminant information due to dependences between X_k and the *entire* set of X_j, $j < k$, (20) restricts this measure to dependences between X_k and subset $\tilde{\mathbf{C}}_{i,k}$. Hence, (20) requires density estimates of dimension of at most $l+1$. Since density estimation has exponential complexity on feature space dimension, the complexity difference between (16) and (8) can be very significant if $l \ll D$. To illustrate this, we analyze a simple example.

TABLE 1
Terms of (17) and (15) When $D = 6$ and $l = 2$

k	i	$I(X_k;\tilde{\mathbf{C}}_{i,k}\vert Y) - I(X_k;\tilde{\mathbf{C}}_{i,k})$	$I(X_k;\tilde{\mathbf{C}}_{i,k}\vert\mathbf{C}_1,\ldots,\mathbf{C}_{i-1}) - I(X_k;\tilde{\mathbf{C}}_{i,k})$
2	1	$I(X_2;X_1\vert Y) - I(X_2;X_1)$	$I(X_2;X_1) - I(X_2;X_1) = 0$
3	1	$I(X_3;\mathbf{C}_1\vert Y) - I(X_3;\mathbf{C}_1)$	$I(X_3;\mathbf{C}_1) - I(X_3;\mathbf{C}_1) = 0$
4	1	$I(X_4;\mathbf{C}_1\vert Y) - I(X_4;\mathbf{C}_1)$	$I(X_4;\mathbf{C}_1) - I(X_4;\mathbf{C}_1) = 0$
4	**2**	$I(X_4;X_3\vert Y) - I(X_4;X_3)$	$I(X_4;X_3\vert\mathbf{C}_1) - I(X_4;X_3)$
5	1	$I(X_5;\mathbf{C}_1\vert Y) - I(X_5;\mathbf{C}_1)$	$I(X_5;\mathbf{C}_1) - I(X_5;\mathbf{C}_1) = 0$
5	**2**	$I(X_5;\mathbf{C}_2\vert Y) - I(X_5;\mathbf{C}_2)$	$I(X_5;\mathbf{C}_2\vert\mathbf{C}_1) - I(X_5;\mathbf{C}_2)$
6	1	$I(X_6;\mathbf{C}_1\vert Y) - I(X_6;\mathbf{C}_1)$	$I(X_6;\mathbf{C}_1) - I(X_6;\mathbf{C}_1) = 0$
6	**2**	$I(X_6;\mathbf{C}_2\vert Y) - I(X_6;\mathbf{C}_2)$	$I(X_6;\mathbf{C}_2\vert\mathbf{C}_1) - I(X_6;\mathbf{C}_2)$
6	**3**	$I(X_6;X_5\vert Y) - I(X_6;X_5)$	$I(X_6;X_5\vert\mathbf{C}_1,\mathbf{C}_2) - I(X_6;X_5)$

Example 1. Let $D = 6$ and $l = 2$. Then, $\mathbf{C}_1 = \{X_1, X_2\}$, $\mathbf{C}_2 = \{X_3, X_4\}$, $\mathbf{C}_3 = \{X_5, X_6\}$, and $C_{\mathcal{C}_l}(\mathbf{X};Y)$ is the sum of the terms in the third column in Table 1. These terms measure discriminant information due to dependences within $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$, (zeroth-order components) and between X_3 and $\mathbf{C}_1$, X_4 and $\mathbf{C}_1$, X_5 and $\mathbf{C}_1$, X_5 and $\mathbf{C}_2$, X_6 and $\mathbf{C}_1$, and X_6 and $\mathbf{C}_2$ (first order). Hence, (16) requires joint density estimates of up to three features. On the other hand, (8) requires densities of up to six features and is *three orders of magnitude* more complex.

4.2 Decompositions for Low-Complexity Feature Selection

Theorem 1 only holds for the decomposition of $\mathbf{X}$ according to (12) and (13). This is not sufficient for feature selection algorithms, which usually evaluate the MI of various subsets of $\mathbf{X}$. For this, the theorem must be expanded to *all* possible feature subsets of $\mathbf{X}$. The extension of the necessary and sufficient condition of (18) to all such subsets is denoted as l-decomposability.

Definition 3. *A feature set* $\mathbf{X}$ *is* l-decomposable *or is decomposable at order* l *if and only if*

$$CI(\mathbf{W};\mathcal{C}_l) = \sum_{k=2}^{|\mathbf{W}|}\sum_{i=1}^{\lceil k-1/l\rceil}\left[I(W_k;\tilde{\mathbf{C}}_{i,k}|\mathbf{C}_1^{i-1},Y) - I(W_k;\tilde{\mathbf{C}}_{i,k}|Y)\right], \quad \forall \mathbf{W} \in \mathcal{S}(\mathbf{X}), \quad (21)$$

where $\mathcal{C}_l$ *and* $\tilde{\mathbf{C}}_{i,k}$ *are built from* $\mathbf{W}$*, as in (12) and (13), and* $\mathcal{S}(\mathbf{X})$ *is the set of all subsets of* $\mathbf{X}$*.*

Since (18) holds for any feature subset $\mathbf{W}$ of an l-decomposable set $\mathbf{X}$, simple application of Theorem 1 shows that the same is true for (16).

Corollary 1. *Let* $\mathbf{X}$ *be an* l-decomposable *feature set,* $\mathbf{W}$ *a subset of* $\mathbf{X}$*, and* $\mathcal{C}_l$ *be a collection of disjoint subsets* $\mathbf{C}_i$ *of cardinality* l *built from* $\mathbf{W}$*, as in (12) and (13). Then*

$$I(\mathbf{W};Y) = M(\mathbf{W};Y) + C_{\mathcal{C}_l}(\mathbf{W};Y), \quad (22)$$

with

$$M(\mathbf{W};Y) = \sum_{k=1}^{|\mathbf{W}|} I(W_k;Y), \quad (23)$$

$$C_{\mathcal{C}_l}(\mathbf{W};Y) = \sum_{k=2}^{|\mathbf{W}|} \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(W_k;\tilde{\mathbf{C}}_{i,k}|Y) - I(W_k;\tilde{\mathbf{C}}_{i,k}) \right], \quad (24)$$

where $\tilde{\mathbf{C}}_{i,k}$ is the subset of features in $\mathbf{C}_i$ whose index is smaller than k, and $\mathbf{C}_1^{i-1} = (\mathbf{C}_1, \ldots, \mathbf{C}_{i-1})$.

Hence, for an l-decomposable set, it is equivalent to adopt (2) or (16) as feature selection cost.

Corollary 2. *If $\mathbf{X}$ is l-decomposable, then the solution of (1) is identical to that of*

$$\mathcal{X}^* = \arg\max_{\mathcal{X}\in\mathcal{S}} \left\{ \sum_k I(X_k;Y) + \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k;\tilde{\mathbf{C}}_{i,k}|Y) - I(X_k;\tilde{\mathbf{C}}_{i,k}) \right] \right\}. \quad (25)$$

In summary, *the infomax subset of an l-decomposable $\mathbf{X}$ can be computed with density estimates of dimension $l+1$.* When $l = D$, there is only one possibility for $\mathcal{C}_l$, namely, $\mathcal{C}_l = \{\mathbf{X}\}$, and (16) is equal to (8). Hence, all feature sets are at least D-decomposable, and in the worst case, feature selection has exponential complexity in the cardinality of $\mathbf{X}$. However, depending on the decomposability order of $\mathbf{X}$, this bound may be very loose. *The intrinsic complexity of feature selection is determined by the decomposability order l of the feature set and not its cardinality.*

5 Low-Complexity Infomax Feature Selection Algorithms

In this section, we derive a family of infomax feature selection algorithms based on the theoretical characterization above.

5.1 A New Family of Algorithms

When $\mathbf{X}$ is l-decomposable, the infomax space is given by (25). When l-decomposability does not hold, (25) provides a *low-complexity approximation* to the optimal solution. In this case, l is denoted as the *order of the approximation*, and we refer to the true decomposability order as l^*. Since all feature sets are (at least) D-decomposable, the optimal solution can always be attained if (25) is solved for all values of l. This suggests 1) developing a family of algorithms parameterized by l, 2) solving the feature selection problem for all l, and 3) retaining the best solution. Note that, given l, (25) can be solved by existing feature selection strategies. In our implementation, we use the popular (greedy) strategy of sequential forward feature selection [66], which leads[2] to Algorithm 1. The MIs of (26) are computed with histograms. When b histogram bins are used per feature, the algorithm can be implemented in $O[D(b^l/l)N^2]$ time. Since N is usually small, the complexity is dominated by b and l, increasing exponentially with the latter.

2. It is worth stressing that the algorithm does not guarantee the best approximation for any l, since the greedy selection of a feature limits the feature groupings of subsequent steps. This is a known limitation of sequential forward selection, e.g., shared by all algorithms in Section 2. It can sometimes be circumvented with heuristics such as floating search [66], [73], [74].

Algorithm 1 (approximate infomax of order l).
Input: feature set $\mathbf{X} = \{X_1, \ldots, X_D\}$, order l, and target number of features N.
set $\mathbf{X}^* = \mathbf{C}_1 = \{X_1^*\}$, where $X_1^* = \arg\max_{X_k\in\mathbf{X}} I(X_k;Y)$, $k = 2$, and $i = 1$.
repeat
  for $X_r \notin \mathbf{X}^*$ **do**

$$\delta_r = I(X_r;Y) + \sum_{p=1}^{\lceil k-1/l \rceil} \left[I(X_r;\tilde{\mathbf{C}}_{p,k}|Y) - I(X_r;\tilde{\mathbf{C}}_{p,k}) \right], \quad (26)$$

  end for
  let $r^* = \arg\max_r \delta_r$.
  if $k-1$ is not a multiple of l **then**
    let $\mathbf{C}_i = \mathbf{C}_i \cup \mathbf{X}_{r^*}$,
  else
    set $i = i+1$, $\mathbf{C}_i = \mathbf{X}_{r^*}$.
  end if
  set $\mathbf{X}^* = \cup_i \mathbf{C}_i$, $k = k+1$,
until $k = N$
Output: $\mathbf{X}^*$.

5.2 Comparison to Other Infomax Methods

The main novelty of Algorithm 1 is the use of (26) as a sequential feature selection rule. In addition to the theoretical motivation above, this rule is interesting in two ways. First, it has an intuitive interpretation: it favors features of 1) large MIs with the class label, 2) low MIs with previously selected features, and 3) large MIs with those features given image class. This enforces three principles that are always at play in feature selection:

1. *Discrimination*. Each selected feature must be as discriminant as possible.
2. *Diversity.* The selected features must not be redundant.
3. *Reinforcement.* Unless, this redundancy is itself discriminant.

Second, it unifies many algorithms previously proposed for information-theoretic feature selection.

In fact, *the first three classes in Section 2 are special cases of the family now proposed*. Methods in the first class, namely, marginal infomax, only use the first term of (26). Slightly abusing the notation, we refer to this as the approximate infomax algorithm of order 0. It enforces the principle of discrimination but not diversity or reinforcement and does not guarantee a compact representation: exactly identical features are selected in consecutive steps, wasting some of the available dimensions. The second and third classes are approximations to (26), with $l = 1$, in which case (26) can be written as

$$I(X;Y) + \sum_{i=1}^{k-1} \left[I(X;X_i^*|Y) - I(X;X_i^*) \right]. \quad (27)$$

Algorithms in the second class, based on (4), simply discard the terms that account for the discriminant power of feature dependencies $(I(X;X_i^*|Y))$, failing to enforce the principle of reinforcement. This can be an overkill, since discriminant dependences can be crucial for fine discrimination between otherwise similar classes. On the other hand, by relying on

TABLE 2
Possible Alternatives to the Cost of (27), Their Relation to the Literature, and Performance (Average and Standard Deviation of Precision-Recall Area (PRA)) on Experiments in Section 7

	Cost	feature selection method	PRA
$\Delta(l=0)$	$I(X;Y)$	Marginal infomax	49.7 ± 15.1
$\Delta(l=1)$	$I(X;Y)+\sum_{i=1}^{k-1}[I(X;X_i^*\|Y)-I(X;X_i^*)]$	approximate infomax order 1	$\mathbf{56.3} \pm 17.7$
$\Delta(l=2)$	$I(X;Y)+\sum_{i=1}^{\lceil (k-1)/2 \rceil}[I(X;\tilde{\mathbf{C}}_{i,k}\|Y)-I(X;\tilde{\mathbf{C}}_{i,k})]$	approximate infomax order 2	55.2 ± 17.0
$\delta_{\min}$	$I(X;Y)+\min_i [I(X;X_i^*\|Y)-I(X;X_i^*)]$	method of [58]	54.1 ± 17.6
δ_{med}	$I(X;Y)+median_i [I(X;X_i^*\|Y)-I(X;X_i^*)]$		52.9 ± 16.9
$\delta_{\max}$	$I(X;Y)+\max_i [I(X;X_i^*\|Y)-I(X;X_i^*)]$		52.6 ± 18.3
$\alpha_{\min}$	$I(X;Y)+\min_i I(X;X_i^*\|Y)$		49.0 ± 13.5
$\beta_{\min}$	$I(X;Y)-\max_i I(X;X_i^*)$		53.5 ± 17.1
β_{avg}	$I(X;Y)-\frac{1}{k-1}\sum_{i=1}^{k-1} I(X;X_i^*)$	mRMR method of [48]	53.4 ± 15.6
α	$I(X;Y)+\sum_{i=1}^{k-1} I(X;X_i^*\|Y)$		50.2 ± 16.9
β	$I(X;Y)-\sum_{i=1}^{k-1} I(X;X_i^*)$	method of [54] with $\xi=1$	53.4 ± 15.7

(7), the algorithms in the third class approximate the summation of (27) by its smallest term.

The excellent empirical performance [48], [58] of algorithms in the second and third classes suggests two hypotheses. The first is that the infomax approximation of first order $(l=1)$ is sufficient for many problems of practical interest. The second is that, even for this approximation, many terms of (27) are neglectable. It is, nevertheless, puzzling that excellent results have been achieved with two very different approximations: the average MI between features (the max-relevance min-redundancy (mRMR) method [48]) and the minimum of the differential MI terms [58]. It is also unclear why these would be the only sensible simplifications. Given that both the minimum differential term and the average of the negative terms perform well, why not consider the smallest among the negative terms, their sum (as proposed in [47], [48], [51], [52], [54], and [55]), or the median of the differential terms? Table 2 presents a number of such alternatives to (27), as well as their empirical performance on a set of experiments to be discussed in Section 7.

6 Image Statistics and Low-Decomposability Order

In this section, we develop an indirect procedure for validating the hypothesis that bandpass features extracted from natural images have low decomposability order.

6.1 l-Decomposability and Image Statistics

From Definition 3, $\mathbf{X}$ is l-decomposable if the conjunctive interference (with respect to subsets of cardinality l) within any of its subsets $\mathbf{W} \subset \mathbf{X}$ is nondiscriminant. This can be illustrated by returning to Example 1, for which the terms of (15) are the entries in the fourth column in Table 1. Note that the nontrivially zero entries (identified by boldface k and i) measure how the dependences in $\mathbf{C_2}$ are affected by $\mathbf{C}_1$ $(k=4, i=2)$, how the dependences in $X_5 \cup \mathbf{C}_2$ are affected by $\mathbf{C}_1$ $(k=5, i=2)$, how the dependences in $X_6 \cup \mathbf{C}_2$ are affected by $\mathbf{C}_1$ $(k=6, i=2)$, and how the dependences in $\mathbf{C_3}$ are affected by $\mathbf{C}_1 \cup \mathbf{C}_2$ $(k=6, i=3)$. $CI(\mathbf{X};\mathcal{C}_l)$ is the sum of these measures and, for l-decomposability to hold, must not be affected by knowledge of the class Y.

In addition to this, l-decomposability requires (18) to hold for any subset $\mathbf{W} \subset \mathbf{X}$. For example, $\mathbf{W}=(X_1, X_3, X_5, X_6)$ produces a table similar to Table 1, with a single nontrivially zero entry $I(X_6; X_5|X_1, X_3) - I(X_6; X_5)$. l-decomposability requires that the interference of (X_1, X_3) on the dependence between X_5 and X_6 be nondiscriminant. Other subsets of the four features give rise to similar constraints on the interference between feature pairs. Hence, in this example, l-decomposability requires all pairwise interferences to be nondiscriminant.

In general, l-decomposability holds if and only if the conjunctive interference (with respect to subsets of cardinality l) within any subset $\mathbf{W}$ of $\mathbf{X}$ is not affected by knowledge of the class label Y. As in Fig. 2, *this does not mean that conjunctive interference is nonexistent but simply that it does not change across classes*. Overall, the sufficient condition for l-decomposability is similar to the sufficient condition for the optimality of marginal infomax. While, in that case, image statistics must satisfy $C(\mathbf{X};Y)=0$, i.e., that *no* dependences in $\mathbf{X}$ are discriminant, in this case, the constraints only affect second-order *subset* dependences: l-decomposability *does not impose constraints on subset dependencies of zeroth or first order, nor does it impose that there are no second-order subset dependences. It only requires these dependences to be such that the conjunctive interference $CI(\mathbf{X};\mathcal{C}_l)$ is nondiscriminant*. This is much less restrictive than what is required for the optimality of marginal infomax. As in that case, the consistency of the statistics in Fig. 1a suggests that for natural images, the hypothesis that l-decomposability holds for small l is not unreasonable. We next turn to the problem of determining this value.

6.2 Indirect Validation of the Low-Order Decomposability Hypothesis

If $\mathbf{X}$ is l^*-decomposable, the infomax feature set can be found with (25) by using $l=l^*$. For approximation orders $l \neq l^*$, the problems of (25) make looser assumptions about feature

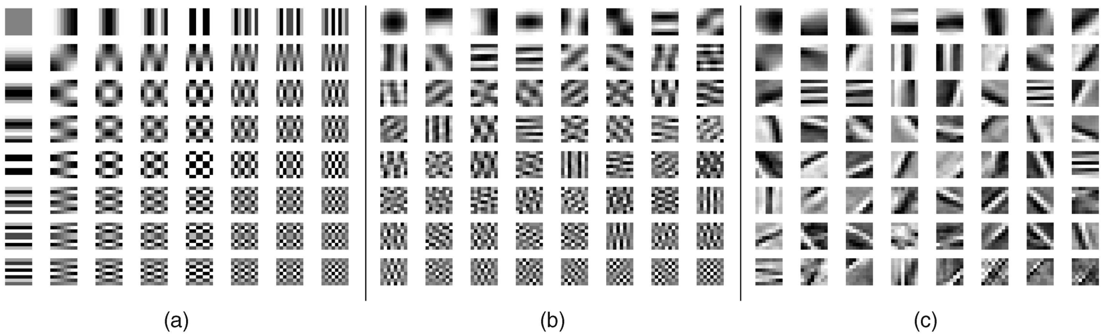

Fig. 3. Basis functions for (a) DCT, (b) PCA, and (c) ICA.

dependences as l increases. $l = 0$ assumes that no feature dependences are discriminant, $l = 1$ assumes that only dependences within feature pairs are important, and so forth, up to $l = D$, where all dependences are accounted for. The decomposability order of the feature set can be determined with recourse to the indirect validation paradigm: the error of classifiers designed on the spaces produced by (25) is expected to decrease with l, leveling off at $l = l^*$. If this produces a consistent estimate of l^* across a number of classification problems, there is strong empirical evidence that $\mathbf{X}$ is l^*-decomposable. If this is repeatedly observed for transformations in a certain class, e.g., wavelets, there is strong evidence that all feature sets in the class are l^*-decomposable.

7 Experiments

In this work, we hypothesize that transformations into sets of bandpass frequency coefficients have low decomposability order. We rely on indirect validation to test this hypothesis.

7.1 Experimental Protocol

All experiments were performed with the Brodatz and Corel image databases. Brodatz is a standard benchmark for texture classification under controlled imaging conditions and with no distractors. Corel is a standard evaluation set for recognition from unconstrained scenes (e.g., no control over lighting or object pose and cluttered backgrounds). Brodatz contains sets of nine patches from 112 gray-scale textures, with a total of 1,008 images. One patch of each texture was used for testing, and the remaining eight were used for training. From Corel, we selected 15 image classes,[3] each containing 100 color images. Train and test sets were then created by assigning each image to the test set with probability 0.2. Evaluation was based on precision and recall (PR), using the test images as queries to a database containing the training set. The PR curve was summarized by its integral, namely, the PRA. In all experiments, feature vectors were extracted from localized image neighborhoods and classification based on (30) with Gauss mixture class-conditional densities. A Gauss mixture with a fixed number of components was learned for each image (results were qualitatively similar for various values, and we report on eight components), examples were assumed independent in (30), and class priors were uniform. Four transformations were considered:

1. discrete cosine transform (DCT),
2. principal component analysis (PCA),
3. independent component analysis (ICA), and
4. wavelet representation (WAV).

The feature space had $D = 64$ per color channel (three layers of wavelet decomposition and 8×8 image blocks), and the observations were extracted with a sliding window. PCA and ICA were learned from 100,000 random training examples. Fig. 3 compares the basis functions learned on Brodatz with those of the DCT.

7.2 Decomposability Order

The decomposability order of all data sets was studied with the indirect validation paradigm. Because the computational cost is exponential on the approximation order l, it is (at this point in time) only feasible to consider small values of this parameter. We have limited all experiments to the range $0 \leq l \leq 2$. Fig. 4 presents the PRA curves obtained with different l for DCT, PCA, and ICA.[4] The most striking observation is that for *all* databases and transformations, $l = 1$ is superior to $l = 0$, but there is no advantage of $l = 2$ over $l = 1$. The constancy of this result suggests that *all feature sets are 1-decomposable*. To understand this constancy, we analyzed the feature rankings in detail. Fig. 5a presents the top-nine features selected on Brodatz for each transformation and value of l. For $l = 0$, the top features are nearly identical: all have very high frequency and do not appear to capture perceptually interesting image structure. This indicates that marginal statistics are not enough for these problems. The solution obtained with $l = 1$ is superior: not only the features appear to be detectors of perceptually relevant image attributes but the same also holds for their pairwise conjunctions. This is shown in Fig. 5b, which presents the optimal pairwise DCT conjunctions. While individual features detect smooth regions, horizontal and vertical bars, horizontal and vertical edges, horizontal and vertical parallel segments, corners, and rounded spots, the set of conjunctions includes detectors of crosses, T- and

3. "Arabian horses," "auto racing," "owls," "roses," "ski scenes," "religious stained glass," "sunsets and sunrises," "coasts," "divers and diving," "Land of the Pyramids" (pictures of Egypt), "English country gardens," "fireworks," "glaciers and mountains," "Mayan and Aztec ruins," and "oil paintings."

4. Qualitatively identical results were obtained with the wavelet and are omitted for brevity.

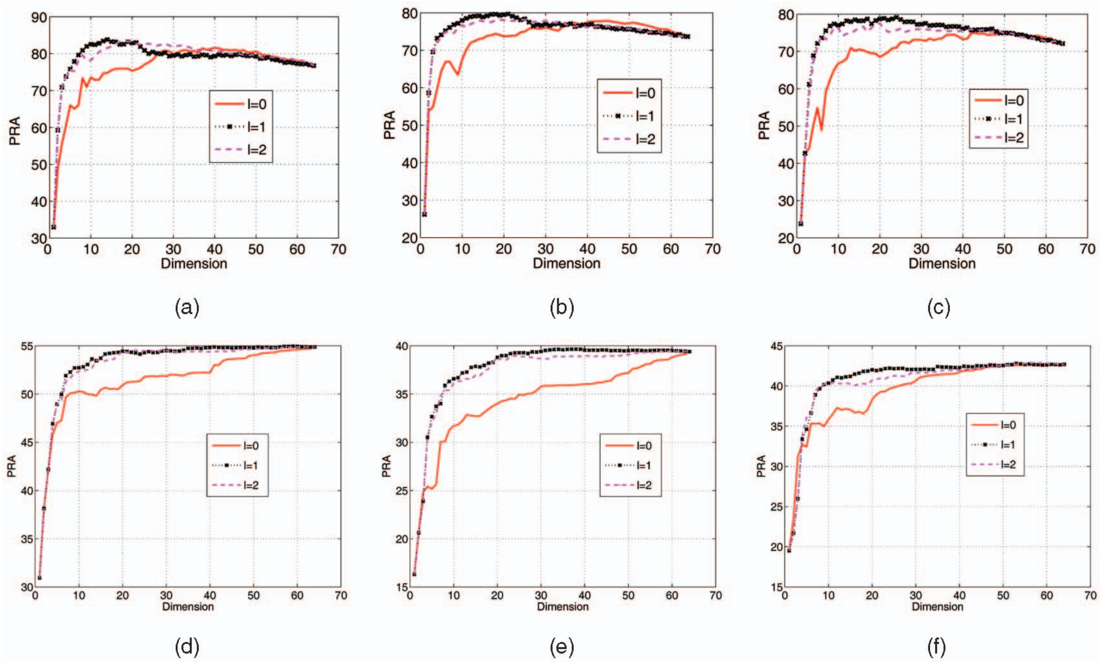

Fig. 4. PRA as a function of the number of features selected by approximate infomax $0 \leq l \leq 2$ for the DCT, PCA, and ICA feature sets on Brodatz and Corel. (a) DCT on Brodatz. (b) PCA on Brodatz. (c) ICA on Brodatz. (d) DCT on Corel. (e) PCA on Corel. (f) ICA on Corel.

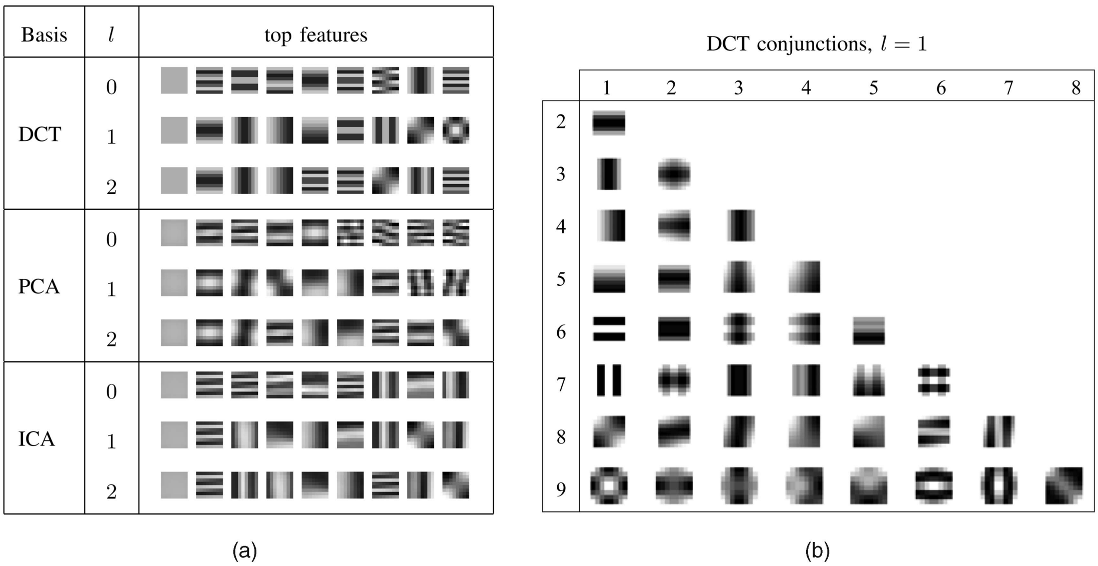

Fig. 5. (a) Top-nine features (in decreasing order, from left to right) selected on Brodatz for the three representations and $0 \leq l \leq 2$. (b) Conjunctions of features that contribute to (16) for the optimal feature set on Brodatz with DCT features and $l = 1$. The basis function at row i and column j of the table was produced by averaging features i and j of the optimal set of (a).

L-junctions, grids, oriented lines, etc.[5] The fact that, for $l = 1$, *features are selected not only by individual discriminant power but also by the discriminant power of pairwise conjunctions makes a significant difference for both classification accuracy* (see Fig. 4) *and perceptual relevance of the visual attributes that they detect* (see Fig. 5). Finally, there is no benefit in considering $l = 2$: both classification performance and perceptual relevance decrease slightly. Because the union of individual features and pairwise conjunctions is very discriminant, the gain of triplets is small. On the other hand, all dimensionality problems (complexity of density estimation and exponential increase in training data requirements) are compounded, and the overall result is a loss.

5. In fact, the set of conjunctions is much larger than that shown. While the table only includes pairwise feature *averages*, the set includes all functions of the same feature pairs.

7.3 Comparison to Previous Methods

The classification performance of (25), with $0 \leq l \leq 2$ (costs $\Delta(l)$), is compared to that of each of the other costs in Table 2. For each transformation and image database, classification performance is summarized by the average PRA of the first N features. N was chosen to guarantee that the number of features needed for optimal performance was available but not a lot more (all methods perform similarly

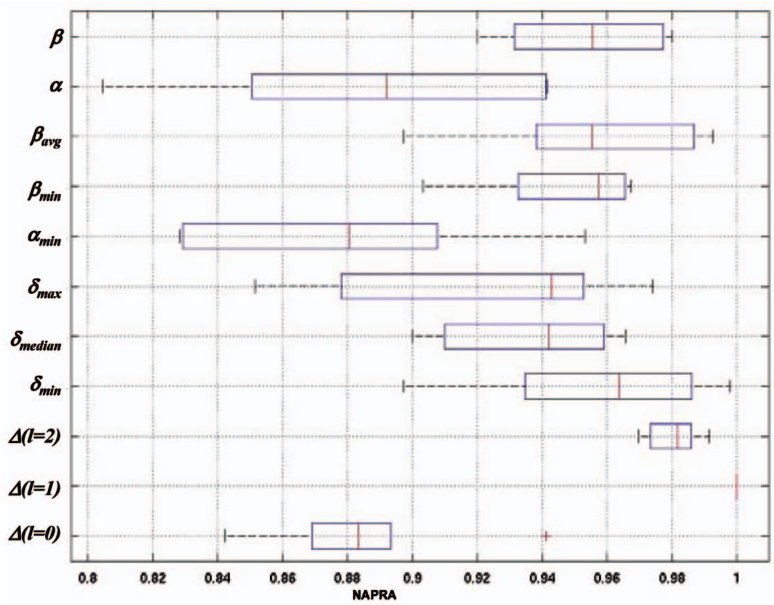

Fig. 6. NAPRA scores for the different costs in Table 2 across feature transforms and databases. Box lines indicate lower, median, and upper quartile values. Dashed lines show the extent of the rest of the data.

when N is close to the total number of features). Using Fig. 4 for guidance, we chose $N = 15$ for Brodatz and $N = 20$ for Corel. Table 2 presents the average and standard deviation of the PRA achieved (across data sets and transforms) by each cost. To facilitate the comparison, we divided (for each data set and transform) the average PRA of each cost by that achieved with $\Delta(l = 1)$. The average of this measure (across data sets and transforms) is denoted as the normalized average PRA (NAPRA) score of the cost. Fig. 6 presents a boxplot of the NAPRA score of each cost across databases and transformations. A number of interesting observations can be made. The first is that $\Delta(l = 1)$ produced the best feature set in *all* cases. The second overall best performer was $\Delta(l = 2)$, followed by the three costs previously proposed in the literature: $\delta_{\min}$ [58], β_{avg} [48], and β [54]. On average, there was no substantial difference between these three costs, although $\delta_{\min}$ performed best. The fact that these are the best approximations to $\Delta(l = 1)$ (among those that we have evaluated) is a possible explanation for their impressive performance in previous experimental comparisons [48], [58]. Of the remaining costs, δ_{median} and $\delta_{\max}$ performed somewhat worse but clearly above the marginal infomax ($\Delta(l = 0)$), while α and $\alpha_{\min}$ did not consistently beat the latter.

Returning to the indirect validation paradigm, these results provide information about the importance, for discrimination, of various aspects of the feature statistics. The first interesting observation is that the *average performance of marginal infomax is close to 90 percent of the best*. This suggests that, for natural images, *most discriminant information is contained in marginal feature statistics*. Given that the marginal infomax is the *only* method that does not require joint density estimates, *it may be the best solution for recognition problems with strict constraints on time or computation*. It is also interesting to investigate which terms of $\Delta(l = 1)$ are responsible for its performance gain over $\Delta(l = 0)$. One observation based on Fig. 6 is that this gain is very nonlinear on the differential terms $\delta_i = I(X; X_i^*|Y) - I(X; X_i^*)$. In particular, the inclusion of a single term, be it the largest ($\delta_{\max}$), median (δ_{median}), or most negative ($\delta_{\min}$), is sufficient to achieve at least half the total gain, with $\delta_{\min}$ achieving 2/3. Hence, while it is important to include one differential term, *the exact choice may not be very important*. This flexibility could be significant when there are complexity constraints. While computing an arbitrary δ_i has linear complexity on the number of features, the search for the best term has quadratic complexity. It follows that *the inclusion of an arbitrary differential term may be a good intermediate solution* (complexity quadratic in histogram bins but linear on features) *between the marginal infomax and an approximate infomax of order 1*. On the other hand, finding the best δ_i [58] requires more computation than evaluating $\Delta(l = 1)$ (due to the search after all terms are computed) and has no advantage.

As an alternative to the differential terms δ_i, Fig. 6 shows that gains can also be obtained by adding terms of each MI type $\alpha_i = I(X; X_i^*|Y)$ and $\beta_i = I(X; X_i^*)$ to $\Delta(l = 0)$. Here, it appears that β_i are much more important than α_i: by themselves, α_i do not even produce a consistent improvement over the marginal infomax. On the other hand, the inclusion of the best β_i (cost $\beta_{\min}$) does not perform nearly as well as the inclusion of the best δ_i (cost $\delta_{\min}$). In fact, the latter performed better than all the α- or β-only approaches considered. Yet, the gains of the β-only costs could, once again, be interesting if there are complexity constraints. Note that, unlike the α terms, they do not depend explicitly on the class Y. They could thus be learned from a generic collection of natural images, *independent of the particular recognition problem under consideration*. In this case, the *complexity of the β costs would be equivalent to that of the marginal infomax*. While it is currently unclear if the performance would remain as in Fig. 6 (where all β_i were estimated from the training sets used for classifier design), this is an interesting topic for further research.

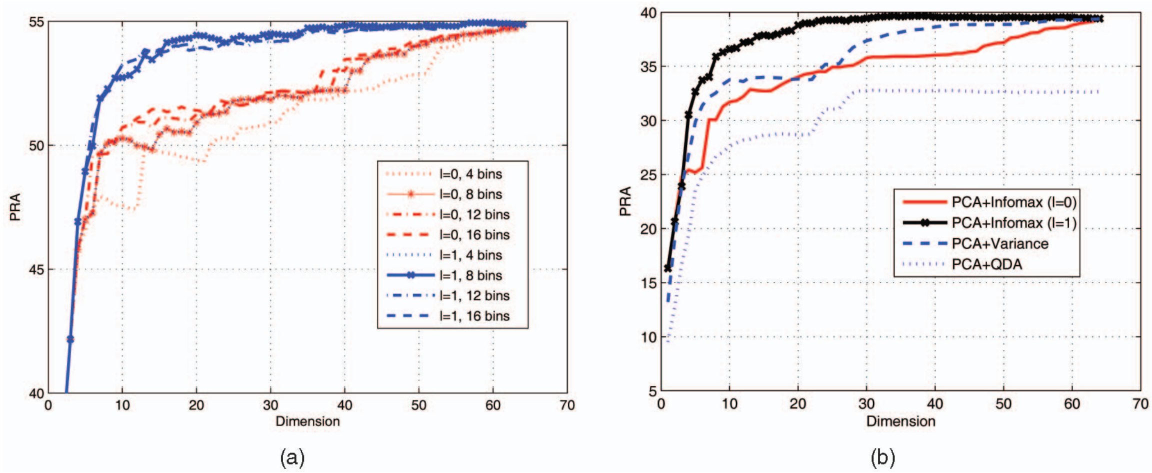

Fig. 7. (a) PRA curves for the DCT on Corel by using $\Delta(l = i)$, $i \in \{0, 1\}$, and various numbers b of histogram bins. (b) Comparison of the PRA curves obtained with infomax and popular methods of equivalent complexity for PCA.

As discussed in Section 2, there is a large literature on β costs, mostly focusing on the role of the parameter ξ of (5) [47], [48], [51], [52], [53], [54], [55]. Fig. 6 suggests that for natural images, this discussion is inconsequent: similar performance was obtained with only one β_i (cost $\beta_{\min}$), their average (cost β_{avg}), or sum (cost β). Different ξ only affected the variance of the NAPRA score, which was smallest for $\xi = 1$. The increased variance of the other weights might explain various sometimes-conflicting claims for their success [47], [48], [51], [52], [54], [55].

In summary, the infomax approximation of order 1 ($\Delta(l = 1)$) outperforms the previous low-complexity methods. It is worth emphasizing that the discussion above is based on the *average* performance of the different costs across data sets and transformations. One important point is that *all previous methods exhibited "breakdown" modes*, i.e., combinations of transformation/database on which their performance was well below average. This can be seen from the limit intervals (dashed lines) in Fig. 6. In almost all cases, the lower bound is close to the average performance of marginal infomax. The only salient exceptions are $\Delta(l = 1)$, which always performed the best, and $\Delta(l = 2)$, which has small variance. These observations suggest that the *main role of the summation in (26) is to assure robustness. While simplifications of this rule can perform well for certain data sets, they compromise generalization*.

7.4 Robustness

Assuming that bandpass transforms are indeed 1-decomposable, we performed some experiments to determine the robustness to the parameter that determines the complexity: the number of histogram bins/axis b (recall that the complexity of approximate infomax of order l is $O(b^l)$). In particular, we repeated the experiment with $l = 0$ and $l = 1$ for values of b in [4, 16]. Fig. 7a presents PRA curves from Corel, with DCT features,[6] showing that recognition accuracy is quite insensitive to this parameter. For both values of l, eight bins are sufficient to achieve accuracy very close to the highest. A loss only occurs for $b = 4$ and, as expected, is more significant for $l = 1$, where the density estimates are two-dimensional.

7.5 Comparison with Scalable Feature Selection Methods

To place the results above in a larger feature selection context, we compared the infomax algorithms with two widely popular methods of *similar complexity*: PCA and its combination with quadratic discriminant analysis [75] (PCA + QDA). Because these methods project all examples onto the PCA subspace, we restricted the comparison to the infomax subset of PCA features. Although PCA is frequently combined with the euclidean or Mahalanobis distances and a nearest neighbor classifier, namely, the popular "Eigenfaces" technique [76], preliminary experiments showed better performance for a Gauss mixture classifier on the PCA subspace. This is identical to the classifier adopted for the infomax features but relies on feature ranking by variance rather than the MI. PCA + QDA is an extension of the popular "Fisherfaces" method [77] and is equivalent to (30) when the PCA coefficients are Gaussian. It was implemented by fitting a multivariate Gaussian to each training image and using (30) to classify all test images.

Fig. 7b compares, on Corel, the PRA curves of PCA + Variance and PCA + QDA with those previously presented for infomax ($l \in \{0, 1\}$) on the PCA space.[7] PCA + QDA performs significantly worse than all other approaches. This is not surprising, given the strong non-Gaussianity of the distributions in Fig. 1a. With the Gaussian mixture classifier, maximum variance and marginal infomax have similar performance,[8] but infomax with $l = 1$ is substantially better. For example, in the PCA case, energy compaction requires about 30 features to reach the accuracy that infomax ($l = 1$) achieves with only 10. For the DCT, the ratio is even larger, closer to 4/1. Visual inspection of recognition results shows significant improvement for queries from classes that share

6. Similar results were obtained on Brodatz and are omitted.

7. Once again, similar results were obtained on Brodatz and are omitted.

8. While maximum variance is somewhat superior to marginal infomax in Fig. 7b, we have seen no consistent differences between the two criteria across all feature spaces.

visual attributes with other classes in the database (see [78] for examples).

8 Discussion

We have studied the hypothesis that high-order dependences of bandpass features contain little information for image discrimination. The hypothesis was characterized formally by the introduction of the concepts of conjunctive interference and decomposability order and the derivation of necessary and sufficient conditions for the feasibility of low-complexity feature selection in terms of these concepts. It was shown that the intrinsic complexity of feature selection is determined by the decomposability order of the feature set: the infomax subset of an l-decomposable set can be computed with density estimates of dimension $l+1$. A family of (locally) optimal feature selection algorithms was then derived for all levels of complexity, and its performance was characterized in two ways. Theoretically, it was shown that various previous information-theoretic feature selection algorithms are approximations to the ones now derived. Experimentally, the latter were shown to consistently outperform the former for a diverse set of images and feature transformations.

Following the indirect validation paradigm, the new feature selection algorithms were used to objectively test the hypothesis of low decomposability order for natural image features. This has shown that while there is a nonnegligible classification gain in modeling feature dependencies (in all cases, $l=1$ outperformed $l=0$), this gain has diminishing returns. Certainly, the benefits of modeling dependencies between triplets $(l=2)$ over pairs $(l=1)$ are at most marginal. While it is possible that there may be some $l>2$ with substantially better performance than $l=1$, the consistent lack of improvement from $l=1$ to $l=2$ across the imagery and features considered in our experiments suggests that this is unlikely. Unfortunately, limitations in computation and database size currently prevent us from experimenting with $l>2$.

A detailed investigation of the $l=1$ case has shown that when pairwise dependences are modeled, the gains are very nonlinear on the rigor of this modeling. In particular, simple modeling of marginal statistics performs fairly well (within 90 percent of the top performance), and the inclusion of a single pairwise differential term, as proposed in [58], can capture as much as 2/3 of what remains. On the other hand, the simple inclusion of the so-called β terms, as proposed in [47], [48], [51], [52], [54], and [55], can also work well. Since β terms do not depend on the particular classification problem under analysis, they could conceivably be learned from a generic image database. In this case, it should be possible to account for dependences with feature selection algorithms that only require the estimation of marginal densities. This remains an interesting topic for further research. The main benefit of accounting for all terms of order 1 seems to be a significant increase in robustness. While the previously proposed approximations can perform very well in some cases and reasonably well on the average, they have all exhibited "breakdown" modes (combinations of features and image databases where the performance was similar to that of marginal statistics). The large variance of their classification performance could explain previous conflicting claims for the superiority of different approximations [47], [48], [51], [52], [54], [55]. On the other hand, the algorithms now proposed performed very robustly, consistently achieving the best results on all data sets. It would therefore be speculative to 1) propose that some of the terms of the simplified MI of order 1 are more important than others or 2) make generic statements about the details of the dependence structure encoded by these terms.

What can thus be said about the structure of the dependencies of bandpass features extracted from natural images? 1-decomposability means that for natural images, the conjunctive interference between individual features is not discriminant. Or, in other words, *the effect, on the dependence between two features, of observing any other feature is constant across image classes*. This is a significantly more precise statement than the hypothesis that feature dependences are constant across classes, with which we started. Although our analysis is limited to features extracted from natural images, this conclusion also appears sensible for modalities such as audio, speech, or language. For example, in the language context, it would imply that the effect of observing a word on the dependence between two other words is constant across document classes. This simply suggests that second-order dependences between words are determined by language and not the specific document classes. It is a fairly mild constraint on the structure of text, e.g., much milder than the common bag-of-words model. It is interesting to note that some experimental observations similar to the ones that we have reported for images have been made for text categorization. These include reports of successful application of marginal infomax [41], [42] and reports of improved performance by the criterion of (6) [44].

Appendix A

Optimality Criteria for Feature Selection

In the most general sense, the optimal feature space for a classification problem $\mathcal{C}_{\mathcal{X}}$ is

$$\mathcal{X}^* = \arg\min_{\mathcal{X}\in\mathcal{S}} J(\mathcal{C}_{\mathcal{X}}), \tag{28}$$

where $J(\cdot)$ is a cost, and $\mathcal{S}$ is the set of range spaces for the transforms under consideration.

A.1 Minimum Bayes Error Features

One measure of the goodness of $C_{\mathcal{X}}$ is the lowest possible probability of error achievable in $\mathcal{X}$, usually referred to as the *Bayes error* [33]:

$$L^*_{\mathcal{X}} = 1 - E_{\mathbf{x}}\left[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})\right], \tag{29}$$

where $E_{\mathbf{x}}$ is the expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$. It depends only on $\mathcal{X}$ and not the classifier itself, and there is at least one classifier that achieves this bound, the *Bayes* decision rule:

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{30}$$

While it is natural to define $\mathcal{X}^\star$ as the space of *minimum Bayes error*, it has long been known that the resulting optimization can be difficult. For example, sequential feature selection is not easy in this setting, since the $\max(\cdot)$ nonlinearity of (29) makes it impossible to decompose the new cost $(E_{\mathbf{X}_n}[\max_i P_{Y|\mathbf{X}_n}(i|\mathbf{x}_n)])$ as a function of the previous best ($E_{\mathbf{X}_c}[\max_i P_{Y|\mathbf{X}_c}(i|\mathbf{x}_c)]$) and a function of the candidate set $\mathbf{X}_a$, where $\mathbf{X}_c$ is the best current subset, and $\mathbf{X}_n = (\mathbf{X}_a, \mathbf{X}_c)$.

A.2 Infomax Features

The infomax formulation has a number of appealing properties.

Lemma 3. *Let $\langle f(i)\rangle_Y = \sum_i P_Y(i) f(i)$ and let $KL[p\|q] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ be the relative entropy between p and q, with integrals replaced by summations for discrete random variables. The following properties hold for the MI, as defined in (2):*

1. *For any two random vectors $\mathbf{X}$ and $\mathbf{Z}$, $I(\mathbf{X};\mathbf{Z}) \geq 0$ with equality if and only if $\mathbf{X}$ and $\mathbf{Z}$ are statistically independent. Furthermore, $I(\mathbf{X};\mathbf{Z}) = KL[P_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})\|P_{\mathbf{X}}(\mathbf{x})P_{\mathbf{Z}}(\mathbf{z})]$.*
2. *$I(\mathbf{X};Y) = \langle KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)\|P_{\mathbf{X}}(\mathbf{x})\right]\rangle_Y$.*
3. *$I(\mathbf{X};Y) = H(Y) - H(Y|\mathbf{X})$, where $H(Y) = -\langle \log P_Y(i)\rangle_Y$ is the entropy of Y, and $H(Y|\mathbf{X}) = -E_{\mathbf{X}}[\langle \log P_{Y|\mathbf{X}}(i|\mathbf{x})\rangle_{Y|\mathbf{X}}]$ is the posterior entropy of Y, given $\mathbf{X}$.*
4. *If $\mathbf{X}_{1,k} = \{X_1, \ldots, X_k\}$, then*

$$I(\mathbf{X}_{1,k};Y) - I(\mathbf{X}_{1,k-1};Y) = I(X_k;Y|\mathbf{X}_{1,k-1}), \tag{31}$$

where

$$I(X;Y|\mathbf{Z}) = \sum_i \int P_{X,Y,\mathbf{Z}}(x,i,\mathbf{z}) \log \frac{P_{X,Y|\mathbf{Z}}(x,i|\mathbf{z})}{P_{X|\mathbf{Z}}(x|\mathbf{z})P_{Y|\mathbf{Z}}(i|\mathbf{z})} dx d\mathbf{z}.$$

Proof. All proofs are either available in [79] or are straightforward consequences of (2). □

Property 1 and the fact that $KL[p\|q]$ is a measure of similarity between distributions p and q show that $I(\mathbf{X};\mathbf{Z})$ is a measure of dependence between $\mathbf{X}$ and $\mathbf{Z}$. For this reason, we frequently refer to $I(\mathbf{X};\mathbf{Z})$ as the *dependence* between $\mathbf{X}$ and $\mathbf{Z}$. Property 2 implies that

$$\mathcal{X}^* = \arg\max_{\mathcal{X}\in\mathcal{S}} \langle KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)\|P_{\mathbf{X}}(\mathbf{x})\right]\rangle_Y, \tag{32}$$

i.e., that infomax feature selection is inherently discriminant: it rewards spaces where the class densities are, on the average, well separated from the mean density. This is a sensible way of quantifying the intuition that optimal discriminant transforms are those that best separate the different classes.

From Property 3, it follows that $\mathcal{X}^* = \arg\min_{\mathcal{X}\in\mathcal{S}} H(Y|\mathbf{X})$. Since entropy is a measure of uncertainty, this implies that the infomax space minimizes the uncertainty about which class is responsible for the observed features. It also establishes a formal connection to the minimization of Bayes error, since in both cases, the optimal space is

$$\mathcal{X}^* = \arg\max_{\mathcal{X}\in\mathcal{S}} E_{\mathbf{X}}\left[\phi\left(P_{Y|\mathbf{X}}(1|\mathbf{X}), \ldots, P_{Y|\mathbf{X}}(M|\mathbf{X})\right)\right],$$

where $\phi(p_1, \ldots, p_M)$ is one of the two functions $\max(p_i)$ and $\langle \log p_i\rangle$, which are both convex and have a number of similar properties (including colocated maxima and minima in the unconstrained probability simplex and interesting relationships between gradients). In fact, there are a number of problems for which the two optimal solutions are identical [62], [80]. Property 4 probably has the greatest practical significance and justifies the adoption of infomax over the minimization of Bayes error. It enables *modular* decompositions of the MI, which are central to the efficient implementation of sequential search and are intuitive. In particular, if $\mathbf{X}^*$ is the current set of selected features, it shows that the feature to be selected at the next step should be

$$X^* = \arg\max_{k|X_k\notin\mathbf{X}^*} I(X_k;Y|\mathbf{X}^*), \tag{33}$$

i.e., the one that most reduces the uncertainty about Y, given $\mathbf{X}^*$. This implies that X^* should 1) be discriminant and 2) have small redundancy with previously selected features.

Appendix B
Proof of Lemma 1

From the chain rule of the MI [79], $I(\mathbf{X},Y) = \sum_{k=1}^D I(X_k;Y|\mathbf{X}_{1,k-1})$. Using the equality

$$\begin{aligned} I(X;Y|\mathbf{Z}) &= E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,Y|\mathbf{Z}}(x,y|\mathbf{z})}{P_{X|\mathbf{Z}}(x|\mathbf{z})P_{Y|\mathbf{Z}}(y|\mathbf{z})}\right] \\ &= E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \right. \\ &\quad \left. + \log \frac{P_{X,Y|\mathbf{Z}}(x,y|\mathbf{z})P_Y(y)}{P_{X,Y}(x,y)P_{Y|\mathbf{Z}}(y|\mathbf{z})} + \log \frac{P_X(x)}{P_{X|\mathbf{Z}}(x|\mathbf{z})}\right] \\ &= I(X;Y) + E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X|Y,\mathbf{Z}}(x|y,\mathbf{z})}{P_{X|Y}(x|y)}\right] - I(X;\mathbf{Z}) \\ &= I(X;Y) + E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,\mathbf{Z}|Y}(x,\mathbf{z}|y)}{P_{X|Y}(x|y)P_{\mathbf{Z}|Y}(\mathbf{z}|y)}\right] \\ &\quad - I(X;\mathbf{Z}) = I(X;Y) + I(X;\mathbf{Z}|Y) - I(X;\mathbf{Z}), \end{aligned} \tag{34}$$

with $X = X_k$ and $\mathbf{Z} = \mathbf{X}_{1,k-1}$, leads to

$$I(\mathbf{X},Y) = \sum_{k=1}^D I(X_k;Y) - \sum_{k=2}^D \left[I(X_k;\mathbf{X}_{1,k-1}) - I(X_k;\mathbf{X}_{1,k-1}|Y)\right],$$

and the lemma follows. □

Appendix C
Proof of Lemma 2

By recursive application of the chain rule of the MI

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

where $\mathbf{C}_1^k = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$. Similarly

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

The lemma follows from (9). □

Appendix D
Proof of Theorem 1

Combining Lemmas 1 and 2

$$\begin{aligned} I(\mathbf{X}; Y) &= \sum_{k=1}^{D} I(X_k; Y) + \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}, Y) \\ &\quad - I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1})\big] = \sum_{k=1}^{D} I(X_k; Y) \\ &\quad - \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} [I(X_k; \tilde{\mathbf{C}}_{i,k}) - I(X_k; \tilde{\mathbf{C}}_{i,k} | Y)] \\ &\quad + \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}, Y) - I(X_k; \tilde{\mathbf{C}}_{i,k} | Y)\big] \\ &\quad - \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}) - I(X_k; \tilde{\mathbf{C}}_{i,k})\big], \end{aligned}$$

it follows that

$$\begin{aligned} I(\mathbf{X}; Y) &= \sum_{k=1}^{D} I(X_k; Y) \\ &\quad + \sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | Y) - I(X_k; \tilde{\mathbf{C}}_{i,k})\big] \end{aligned}$$

if and only if

$$\begin{aligned} &\sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}) - I(X_k; \tilde{\mathbf{C}}_{i,k})\big] = \\ &\sum_{k=2}^{D} \sum_{i=1}^{\lceil k-1/l \rceil} \big[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1^{i-1}, Y) - I(X_k; \tilde{\mathbf{C}}_{i,k} | Y)\big], \end{aligned}$$

and the theorem follows from the definition of $CI(\mathbf{X}; \mathcal{C}_l)$ in (15). □

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments, which substantially improved clarity of the presentation. This work was funded by the US National Science Foundation (NSF) under Awards IIS-0448609 and -0534985.

References

[1] R. Clarke, *Transform Coding of Images.* Academic Press, 1985.
[2] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 674-693, July 1989.
[3] D. Ruderman, "The Statistics of Natural Images," *Network: Computation in Neural Systems,* vol. 5, no. 4, pp. 517-548, 1994.
[4] D. Field, "Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells," *J. Optical Soc. Am. A,* vol. 4, no. 12, pp. 2379-2394, 1987.
[5] D. Field, "What Is the Goal of Sensory Coding," *Neural Computation,* vol. 6, no. 4, pp. 559-601, Jan. 1989.
[6] R. Bucccigrossi and E. Simoncelli, "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Trans. Image Processing,* vol. 8, pp. 1688-1701, 1999.
[7] J. Huang and D. Mumford, "Statistics of Natural Images and Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 1999.
[8] E. Simoncelli and B. Olshausen, "Natural Image Statistics and Neural Representation," *Ann. Rev. of Neuroscience,* vol. 24, pp. 1193-1216, 2001.
[9] A. Torralba and A. Oliva, "Depth Estimation from Image Structure," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 9, pp. 1226-1238, Sept. 2002.
[10] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On Advances in Statistical Modeling of Natural Images," *J. Math. Imaging and Vision,* vol. 18, pp. 17-33, 2003.
[11] F. Long and D. Purves, "Natural Scene Statistics as the Universal Basis of Color Context Effects," *Proc. Nat'l Academy of Sciences USA,* vol. 100, no. 25, pp. 15190-15193, 2003.
[12] B. Olshausen and D. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature,* vol. 381, pp. 607-609, 1996.
[13] A. Bell and T. Sejnowski, "The Independent Components of Natural Scenes Are Edge Filters," *Vision Research,* vol. 37, no. 23, pp. 3327-3328, Dec. 1997.
[14] J.H. van Hateren and D.L. Ruderman, "Independent Component Analysis of Natural Image Sequences Yields Spatiotemporal Filters Similar to Simple Cells in Primary Visual Cortex," *Proc. Royal Soc. B,* vol. 265, pp. 2315-2320, 1998.
[15] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image Denoising Using Scale Mixtures of Gaussians in the Wavelet Domain," *IEEE Trans. Image Processing,* vol. 12, no. 11, pp. 1338-1351, Nov. 2003.
[16] P. Moulin and L. Juan, "Analysis of Multiresolution Image Denoising Schemes Using Generalized Gaussian and Complexity Priors," *IEEE Trans. Information Theory,* vol. 45, pp. 909-919, Apr. 1999.
[17] A. Levin, A. Zomet, and Y. Weiss, "Learning How to Inpaint from Global Image Statistics," *Proc. Ninth IEEE Int'l Conf. Computer Vision (ICCV '03),* pp. 305-312, 2003.
[18] S. Roth and M. Black, "Fields of Experts: A Framework for Learning Image Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '05),* vol. 2, pp. 860-867, 2005.
[19] N. Farvardin and J. Modestino, "Optimum Quantizer Performance for a Class of Non-Gaussian Memoryless Sources," *IEEE Trans. Information Theory,* May 1984.
[20] Y. Weiss, "Deriving Intrinsic Images from Image Sequences," *Proc. Seventh IEEE Int'l Conf. Computer Vision (ICCV '01),* vol. 2, pp. 68-75, 2001.
[21] M. Do and M. Vetterli, "Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance," *IEEE Trans. Image Processing,* vol. 11, no. 2, pp. 146-158, Feb. 2002.
[22] S. Chang, B. Yu, and M. Vetterli, "Adaptive Wavelet Thresholding for Image Denoising and Compression," *IEEE Trans. Image Processing,* vol. 9, no. 9, pp. 1532-1546, Sept. 2000.
[23] M. Heiler and C. Schnorr, "Natural Image Statistics for Natural Image Segmentation," *Int'l J. Computer Vision,* vol. 63, no. 1, pp. 5-19, 2005.
[24] F. Attneave, "Informational Aspects of Visual Perception," *Psychological Rev.,* vol. 61, pp. 183-193, 1954.
[25] H. Barlow, "The Coding of Sensory Messages," *Current Problems in Animal Behaviour,* W. Thorpe and O. Zangwill, eds., pp. 331-360, Cambridge Univ. Press, 1961.
[26] H. Barlow, "Redundancy Reduction Revisited," *Network: Computation in Neural Systems,* vol. 12, pp. 241-253, 2001.
[27] B. Olshausen and D. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1," *Vision Research,* vol. 37, pp. 3311-3325, 1997.
[28] O. Schwartz and E. Simoncelli, "Natural Signal Statistics and Sensory Gain Control," *Nature Neuroscience,* vol. 4, pp. 819-825, 2001.
[29] S. Deneve, P. Latham, and A. Pouget, "Reading Population Codes: A Neural Implementation of Ideal Observers," *Nature Neuroscience,* vol. 2, pp. 740-745, 1999.

[30] A. Pouget, P. Dayan, and R. Zemel, "Information Processing with Population Codes," *Nature Reviews Neuroscience,* vol. 1, no. 2, pp. 125-132, 2000.

[31] D. Gao and N. Vasconcelos, "Discriminant Saliency for Visual Recognition from Cluttered Scenes," *Advances in Neural Information Processing Systems 17,* L.K. Saul, Y. Weiss, and L. Bottou, eds., pp. 481-488, 2005.

[32] A. Levin, A. Zomet, and Y. Weiss, "Separating Reflections from a Single Image Using Local Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '04),* vol. 1, 2004.

[33] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley & Sons, 2001.

[34] N. Vasconcelos and G. Carneiro, "What Is the Role of Independence for Visual Recognition?" *Proc. Seventh European Conf. Computer Vision (ECCV),* 2002.

[35] A. Treisman and G. Galade, "A Feature Integration Theory of Attention," *Cognitive Psychology,* vol. 12, no. 1, pp. 97-136, 1980.

[36] A. Treisman and S. Gormican, "Feature Analysis in Early Vision: Evidence from Search Asymmetries," *Psychological Rev.,* vol. 95, no. 1, pp. 15-48, 1988.

[37] A. Treisman and S. Sato, "Conjunction Search Revisited," *J. Experimental Perception and Performance,* vol. 16, pp. 459-478, 1990.

[38] K. Cave and J. Wolfe, "Modeling the Role of Parallel Processing in Visual Search," *Cognitive Psychology,* vol. 22, no. 5, pp. 225-271, 1990.

[39] J. Wolfe, "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bull. and Rev.,* vol. 1, no. 2, pp. 202-238, 1994.

[40] J. Wolfe and T. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nature Reviews Neuroscience,* vol. 5, pp. 495-501, 2004.

[41] D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," *Proc. Workshop Speech and Natural Language,* pp. 212-217, 1992.

[42] S. Dumais and H. Chen, "Hierarchical Classification of Web Content," *Proc. ACM SIGIR '00,* pp. 256-263, 2000.

[43] S. Dumais, "Using SVMs for Text Categorization," *IEEE Intelligent Systems,* vol. 13, no. 4, pp. 21-23, 1998.

[44] G.W.F. Lochovsky and Q. Yang, "Feature Selection with Conditional Mutual Information Maximin in Text Categorization," *Proc. 13th ACM Conf. Information and Knowledge Management (CIKM '04),* pp. 342-349, 2004.

[45] Y. Seo, A. Ankolekar, and K. Sycara, "Feature Selection for Extracting Semantically Rich Words," Technical Report CMU-RI-TR-04-18, Robotics Inst., Carnegie Mellon Univ., Mar. 2004.

[46] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," *Proc. 18th Int'l Conf. Machine Learning (ICML '01),* pp. 601-608, 2001.

[47] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Proc. IEEE Bioinformatics Conf. (CSB '03),* pp. 523-528, 2003.

[48] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[49] P. Zarjam, M. Mesbah, and M. Boashash, "An Optimal Feature Set for Seizure Detection Systems for Newborn EEG Signals," *Proc. IEEE Int'l Symp. Circuits and Systems (ISCAS '03),* vol. 5, 2003.

[50] E. Grall-Maes and P. Beauseroy, "Mutual Information-Based Feature Extraction on the Time-Frequency Plane," *IEEE Trans. Signal Processing,* vol. 50, no. 4, pp. 779-790, 2002.

[51] G. Tourassi, E. Frederick, M. Markey, and C. Floyd Jr., "Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis," *Medical Physics,* vol. 28, p. 2394, 2001.

[52] T. Chow and D. Huang, "Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information," *IEEE Trans. Neural Networks,* vol. 16, no. 1, pp. 213-224, 2005.

[53] G. Barrows and J. Sciortino, "A Mutual Information Measure for Feature Selection with Application to Pulse Classification," *Proc. IEEE Int'l Symp. Time-Frequency and Time-Scale Analysis,* 1996.

[54] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks,* vol. 5, no. 4, pp. 537-550, July 1994.

[55] M. Kwak and C. Choi, "Input Feature Selection for Classification Problems," *IEEE Trans. Neural Networks,* vol. 13, no. 1, pp. 143-159, 2002.

[56] P. Scanlon, G. Potamianos, V. Libal, and S. Chu, "Mutual Information Based Visual Feature Selection for Lipreading," *Proc. Int'l Conf. Spoken Language Processing,* pp. 857-860, 2004.

[57] H. Yang and J. Moody, "Data Visualization and Feature Selection: New Algorithms for Non-Gaussian Data," *Proc. Neural Information Processing Systems,* 2000.

[58] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *The J. Machine Learning Research,* vol. 5, pp. 1531-1555, 2004.

[59] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual Features of Intermediate Complexity and Their Use in Classification," *Nature Neuroscience,* vol. 5, no. 7, pp. 1-6, 2002.

[60] M. Vidal-Naquet and S. Ullman, "Object Recognition with Informative Features and Linear Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2003.

[61] F. Jurie and B. Triggs, "Creating Efficient Codebooks for Visual Recognition," *Proc. 11th IEEE Int'l Conf. Computer Vision (ICCV '05),* vol. 1, 2005.

[62] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity," *Neural Information Processing Systems,* 2002.

[63] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity: Optimality and Implications for Visual Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2003.

[64] N. Vasconcelos and M. Vasconcelos, "Scalable Discriminant Feature Selection for Image Retrieval and Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2004.

[65] D. Koller and M. Sahami, "Toward Optimal Feature Selection," *Proc. 13th Int'l Conf. Machine Learning (ICML),* 1996.

[66] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 153-158, Feb. 1997.

[67] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, 1997.

[68] R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics,* vol. 26, no. 5, pp. 1651-1686, 1998.

[69] C. Ratanamahatana and D. Gunopulos, "Feature Selection for the Naive Bayesian Classifier Using Decision Trees," *Applied Artificial Intelligence,* vol. 17, no. 5, pp. 475-487, 2003.

[70] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[71] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts," *Int'l J. Computer Vision,* vol. 56, no. 3, pp. 151-177, 2004.

[72] T. Cover and J.V. Campenhout, "On the Possible Orderings in the Measurement Selection Problem," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 7, no. 9, Sept. 1977.

[73] P. Pudil, J. Novovičová, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters,* vol. 15, no. 11, pp. 1119-1125, 1994.

[74] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical Learning of Multi-View Face Detection," *Proc. Seventh European Conf. Computer Vision (ECCV),* 2002.

[75] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer, 2001.

[76] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, 1991.

[77] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[78] *Comparison of Infomax and Maximum Variance Features,* http://www.svcl.ucsd.edu/projects/infomax/examples.htm, 2008.

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

[80] M. Vasconcelos and N. Vasconcelos, "Some Relationships between Minimum Bayes Error and Information Theoretical Feature Extraction," *Proc. SPIE,* vol. 5807, p. 284, 2005.

Manuela Vasconcelos received the Licenciatura in electrical engineering from the Universidade do Porto, Porto, Portugal, in 1986, the MS degree from the Massachusetts Institute of Technology in 1993, and the PhD degree from Harvard University in 2003. Since 2003, she has been a visiting researcher in the Statistical Visual Computing Laboratory, University of California, San Diego. Her research interests include computer vision.

Nuno Vasconcelos received the Licenciatura in electrical engineering and computer science from the Universidade do Porto, Porto, Portugal, in 1988 and the MS and PhD degrees from the Massachusetts Institute of Technology in 1993 and 2000, respectively. From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which became the HP Cambridge Research Laboratory in 2002. In 2003, he joined the Department of Electrical and Computer Engineering, University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. His research work spans various areas, including computer vision, machine learning, signal processing and compression, and multimedia systems. He is the author of more than 50 peer-reviewed publications. He is the recipient of a US National Science Foundation CAREER Award and a Hellman Fellowship. He is a senior member of the IEEE.