# Pattern Recognition: Historical Perspective and Future Directions

**Azriel Rosenfeld,[1] Harry Wechsler[2]**

[1] Center for Automation Research, University of Maryland, College Park, MD 20742-3275;
Email: ar@cfar.umd.edu

[2] Department of Computer Science, George Mason University, Fairfax, VA 22030-4444;
Email: wechsler@cs.gmu.edu

*Received 19 December 1999; revised 30 March 2000*

> "What being walks sometimes on two feet, sometimes on three, and sometimes on four, and is weakest when it has the most?"
>
> —*The Sphinx's Riddle*

**ABSTRACT:** Pattern recognition is one of the most important functionalities for intelligent behavior and is displayed by both biological and artificial systems. Pattern recognition systems have four major components: data acquisition and collection, feature extraction and representation, similarity detection and pattern classifier design, and performance evaluation. In addition, pattern recognition systems are successful to the extent that they can continuously adapt and learn from examples; the underlying framework for building such systems is predictive learning. The pattern recognition problem is a special case of the more general problem of statistical regression; it seeks an approximating function that minimizes the probability of misclassification. In this framework, data representation requires the specification of a basis set of approximating functions. Classification requires an inductive principle to design and model the classifier and an optimization or learning procedure for classifier parameter estimation. Pattern recognition also involves categorization: making sense of patterns not previously seen. The sections of this paper deal with the categorization and functional approximation problems; the four components of a pattern recognition system; and trends in predictive learning, feature selection using "natural" bases, and the use of mixtures of experts in classification. © 2000 John Wiley & Sons, Inc. Int J Imaging Syst Technol, 11, 101–116, 2000

**Key words:** pattern recognition; feature extraction; representation; functional approximation; categorization; induction; predictive learning; feature selection; classification; performance evaluation

## I. INTRODUCTION

Pattern recognition is one of the most important functionalities for intelligent behavior and it is displayed by both biological and artificial systems. Biological organisms have to recognize specific patterns and respond appropriately for survival. For example, antibodies attack foreign intruders, our ears capture sound and speech, and animals locate edible plants and capture prey. In artificial systems, remote sensing involves the classification of spectral data for ecosystem and land management, optical character readers (OCRs) have to comprehend written text, and biometrics seeks human identity from the way people look—face, iris, and retina—or act—gait, fingerprints, and/or hand geometry. Furthermore, robots are faced with recognizing obstacles, including the layout and identities of surrounding objects for safe navigation and efficient manipulation, and bioengineering involves, for example, the reading and interpretation of electrocardiogram (EKG) charts. When the patterns are of a visual nature, one can regard pattern recognition as supplementary to computer vision, providing the abilities of interpretation and classification. Excellent reference books on pattern recognition include the classic texts of Duda and Hart (1973) and Fukunaga (1972, 1990), and more recently, those of Bishop (1995), Vapnik (1995, 1998), Ripley, (1996), Cherkassky and Mulier (1998), and Duda et al. (2000).

Pattern recognition systems have four major components: data acquisition and collection, feature extraction and representation, similarity detection and pattern classifier design, and performance evaluation. In addition, pattern recognition systems are successful to the extent that they can continuously adapt and learn; they require a flexible memory capable of learning from examples and of similarity-based classification (Poggio, 1990). The underlying framework for building pattern recognition systems is that of predictive learning. The pattern recognition problem is a particular case of the more general problem of statistical regression; it seeks an approximating function that minimizes the probability of misclassification error. In this framework, data representation requires the specification of a set of approximating functions, called a dictionary ("features"), to choose from; an inductive principle to design and model the classifier; and an optimization (learning) procedure for the full-fledged definition of the classifier using proper parameter estimation. The inductive principle is fundamental, as it provides a general prescription for what to do with the training data in order to learn the classifier. Conversely, a learning method is a constructive implementation of an inductive principle, i.e., an optimization or parameter estimation procedure, for a given set of approximating functions

*Correspondence to:* Azriel Rosenfeld

in which some specific classifier model is sought, such as feedforward nets with sigmoid units or decision trees (DTs).

As the last sentence suggests, pattern recognition is closely related to connectionist neural networks. Renewed interest in the early 1980s in connectionist networks, as an alternative to statistical pattern recognition and artificial intelligence (AI), can be attributed to two factors. The first is the realization that an approximating function of sufficient complexity can approximate any (continuous) target function with arbitrary accuracy; the second is the ability to train multilayer and nonlinear networks using backpropagation. An excellent perspective on neural networks is provided by Grossberg (1988). An early attempt at establishing both beneficial relationships and basic differences between (statistical) pattern recognition and neural networks was made by Cherkassky et al. (1994). There is a wide range of opinions on the utility of artificial neural networks (ANNs) for statistical inference. The following differences between the two approaches have been listed by Cherkassky et al. (1994):

1. Goals of modeling: In statistics, the goal is interpretability, which favors structured models; in ANN research, the main objective is generalization/prediction.
2. Model complexity: Usually, although not always, ANNs deal with a large amount of training data (i.e., thousands of samples), whereas statistical methods use much smaller training sets. Hence, ANN models usually have higher complexity (number of parameters or weights) than statistical methods.
3. Batch vs. flow-through processing: Most statistical methods utilize the whole training set (batch mode), whereas ANNs favor iterative processing, known as flow-through methods in statistics. Iterative ANN processing requires many presentations of the data and uses slow computational methods such as gradient descent. Statistical methods are usually much faster.
4. Computational complexity: Because statistical methods extract information from the entire training set, they tend to be more complex and more difficult for nonstatisticians to use. ANN methods tend to be computationally simpler, albeit at the expense of recycling (multiple presentation of) the training data. ANN methods can usually be easily understood and applied by novice users.
5. Robustness: ANNs appear to be more robust than statistical methods with respect to statistical tuning. Confidence intervals are routinely provided in statistical methods but are usually lacking in most ANN application studies, even though there is a growing interest in using them (see Section VIII, "Predictive Learning").

The outline of this survey is as follows. The next two sections consider the categorization and functional approximation problems that arise in pattern recognition and the means and ways to solve them. Each of the following four sections is devoted to one of the major components making up a pattern recognition system: data acquisition, feature extraction, pattern classifiers, and performance evaluation. The next three sections address forthcoming developments and trends in predictive learning, natural bases, and mixtures of experts.

## II. CATEGORIZATION

Pattern recognition tasks include classification or identification (i.e., recognizing a previously seen object as such) and categorization (i.e., making sense of novel shapes not previously seen). As an example, face recognition belongs to identification, whereas gender and ethnicity classification correspond to categorization. (We will often use face recognition as an example in this paper, because it is a problem of great current interest.) The distinction between these two recognition tasks can be traced to categorization requiring a larger degree of generalization than mere recognition. "Barring special (albeit behaviorally important) cases such as face recognition, entry-level (Jolicoeur et al., 1984) names of objects correspond to categories rather than individuals, and it is the category of the object that the visual system is required to determine" (Duvdevani-Bar and Edelman, 1999, p. 203). In the remainder of this paper, recognition and categorization will be used interchangeably, as both of them are ultimately concerned with pattern classification.

A "fundamental question about cognition concerns how knowledge about a category is acquired through encounters with examples of the category" (Knowlton and Squire, 1993, p. 1747). The obvious dichotomy for encoding categories would provide for the corresponding regions of identity to be induced by either specific exemplars or their abstracted representations. Categorization is also characteristic of recall and reasoning, when one realizes that youngsters reason by recall, whereas adults recall by reasoning. Memory-based reasoning, characteristic of the exemplar approach to categorization, supports the view that categorical concepts are implicitly defined in terms of some of the exemplars encountered; it does not call for stored abstractions and/or prototypes (Estes, 1993). This view is shared by Knowlton and Squire, who suggest (p. 1747) that "category-level knowledge has no special status but emerges naturally from item memory," and that a novel probe "would be endorsed as belonging to a particular category as a function of the similarity between the new item and the exemplars of that category already stored in memory."

The abstractive approach, on the other hand, defines regions of identity in terms of prototypes induced through any of the available learning methods. For face recognition, memory-based reasoning defines the face space in terms of face exemplars; this is known as absolute-based coding (ABC). Driven by clustering and density measures, ABC regards typicality, e.g., for ethnicity, as depending on the local density surrounding some exemplar face(s). Sparseness of the exemplar space is to be sought as it leads to a lesser number of misclassification errors. The abstract-competing alternative for face space definition, norm-based coding (NBC), defines typicality in terms of the distance from a prototype that abstracts learned faces and draws from cognitive research on schemata (Neisser, 1967). Recent experiments reported by Rhodes et al. (1998) seem to suggest that the ABC model compares favorably with the NBC model, even though it may not provide a complete account of all known effects related to face recognition.

Explanations of the caricature advantage (caricaturized faces are recognized faster than veridicals) usually involve an encoded norm (NBC) face model (Rhodes et al., 1987). Most recently, further support for the ABC exemplar model came from Lewis and Johnston (1999) who suggested that faces are encoded in a multidimensional Voronoi diagram based on normally distributed face space representations. The Voronoi tessellation model accounts for many of the empirical findings on face recognition and caricaturing without requiring a norm face. The questions that still require an answer for both memory-based and abstract representations are: How are patterns described as (compact and discriminating) categorically self-contained concepts? How is similarity measured and under what metric (Valentine, 1991; Beale and Kiel, 1995; Burton

and Vokey, 1998)? The apparent and unifying solution to these questions is that both models must be represented in terms of some basis whose "natural" dimensions evolve over time in response to the spatiotemporal statistics of the patterns that are encountered. It is most likely that the dimensions for such a representation are abstracted rather than involving physical features.

If one were to accept the NBC explanation for memory encodings, one follow-up question is: Why restrict ourselves to one norm only? Edelman (1999) indeed considers this very question when he advances the concept of a chorus of prototypes. Using Shepard's (1968) notion of second-order isomorphism, Edelman suggests that a pattern, restricted for now to shape only and not accounting for structural relationships, is represented internally by the response of a few tuned modules. Each module is broadly selective for some reference shape, and measures the similarity of that shape to the stimulus. Categorization for Edelman now calls for pattern representation *of* similarity instead of representation *by* similarity and it provides for both generalization and shape constancy. The approach draws from research on regularization networks (Poggio and Girosi, 1990), which are similar in concept to radial basis function (RBF) classifiers (see Section VI) for which the prototypes or modules are the result of $k$-means clustering or estimation-maximization (EM)-like methods (Bishop, 1995).

## III. FUNCTIONAL APPROXIMATION

If what pattern recognition does is mostly about categorization, the way to achieve this is through functional approximation, as it seeks optimal classification (I/O) mappings, i.e., pattern classifiers, within the predictive learning framework. Functional approximation deals with a system characterized by several, possibly many, measurable (observable) quantities, called variables. The variables are divided into two groups. The variables in one (input) group are referred to as independent variables (in applied mathematics), explanatory/predictor variables (in statistics), or input variables (in neural networks/machine learning). The variables in the other (output) group also have different names depending on the field of study: dependent variables (applied mathematics), responses (statistics), or output variables (neural networks/machine learning). The goal is to develop a computational relationship (formula/algorithm) between the inputs and the outputs for determining/predicting/estimating the values of the output variables, given the values of the input variables (Friedman, 1994). Different techniques are available for such functional approximation, including regression and density estimation, corresponding to the input variables being continuous, and classification, corresponding to the variables being discrete/categorical. Another taxonomy for functional approximation consists of supervised vs. unsupervised methods, corresponding to the case where the true output labels are provided, vs. the case when they are not provided and probability density function estimation and/or clustering is required.

Functional approximation is an old problem. Two basic strategies have been employed in the past for solving it. One strategy, characteristic of engineering disciplines, attempts to solve it using first/analytical principles, whereas the other strategy, characteristic of biological systems, employs empirical/adaptive modeling. Functional approximation can be thought of as a relationship $y = f(x) +$ "error," where the error is due to (measurement) noise and possibly also to "unobserved" input variables. The main issues to be addressed are related to prediction (generalization) ability, data and dimensionality reduction (complexity), and explanation/interpretation capability (Cherkassky et al., 1994). An important problem, arising repeatedly in this review, relates to the fact that functional approximation in general, and pattern recognition in particular, can become relevant to real applications only when one realizes that the set of data available for learning the proper mappings is both finite and noisy. To address this problem, specific training strategies, and tradeoffs between observed training errors and complexity-driven confidence intervals, have to be considered using the framework of predictive learning (see Section VIII).

Functional approximation methods can be discussed in terms of the representation scheme used for the target function, the optimization strategy used to derive the target function, and the interpretation capability (Friedman, 1994). The representation scheme usually assumes that the target function is estimated as a linear combination of basis functions ("atoms"), drawn from an appropriate dictionary, and corresponds to a basis function expansion. The availability of an appropriate dictionary should not be taken for granted and can involve much effort. An up and coming trend, that of finding some natural basis for defining the proper dictionary, is discussed in Section IX. We describe the putative role that adaptive and evolutionary methods can play in defining the atoms making up such dictionaries.

There are many possible ways to perform functional approximation in terms of the three dimensions mentioned above, and it is of interest to compare them. If the comparison relates to representation, one is tempted to consider the number of terms involved in the expansion. This criterion is not enough by itself (the bases used can differ in complexity); so, one needs to assess representational complexity in terms of minimum description length (MDL) codes (see also the discussion in Section V on the quality of image reconstruction). As an example, ANNs use fixed (sigmoid) univariate basis functions of linear combinations (projections) of input variables. Projection pursuit regression (PPR), characteristic of statistical methods, uses arbitrary univariate basis functions of such projections. Because PPR employs more complex basis functions than ANN, its estimates generally involve fewer terms than those employed by ANN for the same parameter estimation problem (Cherkassky and Mulier, 1998).

Regarding optimization strategies, pattern recognition usually involves parameter estimation, whereas the closely related (see Section I) neural networks are characteristic of nonparametric estimation methods. Parameter estimation corresponds to nonadaptive methods if a preselected set of basis functions is available and only expansion coefficients are sought (using least-squares approximation). As the form of the target function is assumed known, parameter estimation thus introduces strongly biased assumptions about an unknown target function. Nonparametric methods, on the other hand, seek optimal bases, belong to the class of adaptive methods, and are characteristic of neural networks. Nonparametric methods make no assumptions about the target function and instead consider a family (structure) of approximating functions indexed (ordered) by some complexity parameter. The length description of a network and the VC (Vapnik-Chervonenkis) dimension are examples of complexity parameters used to induce a structure on a network architecture. Adaptive methods involve difficult nonlinear problems and optimization becomes very important. Stepwise selection, where the basis functions are estimated one at a time, is characteristic of backfitting (statistical) methods. Optimization over the whole set of basis functions is characteristic of connectionism.

Both parametric and nonparametric estimation methods can be further classified as global or local. Global methods include linear and polynomial regression; their local counterparts include kernel

smoothers and splines. Local parametric methods are applicable only to low-dimensional problems due to the inherent sparseness of small-sample statistics in high-dimensional spaces. Multilayer and PPR networks are examples of global connectionist methods. Kernel and fuzzy methods are characteristic of local connectionist methods. Generalized memory-based learning is another example of an adaptive local connectionist method and corresponds to case-based reasoning and memory-based learning as employed in artificial intelligence. Note that all learning (estimation) algorithms employ a bias mechanism, referred to as inductive bias, to restrict the hypothesis space, in terms of the target functions under consideration, and/or to rank the functional approximations (hypotheses). Induction is not necessarily truth preserving, as compared to deduction, where truth-preserving operators only expand existing knowledge. As an example, learning by induction, a fundamental incremental and (symbolic) machine learning method, trains over both positive and negative (counter-) examples. It uses generalization and specialization operators to define a minimal version space (Mitchell, 1997b) for concept (category) formation, subject to biases similar to MDL.

Functional approximation theory can be traced back to Weierstrass. The well-known Weierstrass approximation theorem states that for any continuous real-valued function $f$ defined on a closed interval $[a, b]$, and for any given positive constant $e$, there exists a (real-coefficient) polynomial $y$ such that $|y(x) - f(x)| < e$ for every $x$ in $[a, b]$. In other words, every continuous function can be uniformly approximated by a polynomial, and polynomials can thus serve as universal approximators. Several theoretical results, starting with one due to Kolmogorov (1937), have shown that multilayer feedforward networks can also serve as universal approximators. Specifically, Kolmogorov's theorem shows that for any given continuous function $f$: $[0, 1]^n \rightarrow \mathbf{R}^m$, $f$ can be realized by a three-layer feedforward neural network having $n$ fanout processing elements in the first (input) layer, $(2n + 1)$ processing elements in the middle (hidden) layer, and $m$ processing elements in the top (output) layer.

Theoretical rather than constructive, Kolmogorov's result shows that mapping of arbitrary continuous functions from the $n$-dimensional cube $[0, 1]^n$ to the real numbers, **R,** in terms of functions of only one variable, is possible. Similar results regarding the ability of neural networks to serve as universal approximators have been obtained by Cybenko (1989) and by Wang and Mendel (1992) using a fuzzy (systems) framework. Wang and Mendel have considered fuzzy systems represented as a series of expansions of fuzzy basis functions—algebraic superpositions of fuzzy membership functions. They were able to show that linear combinations of fuzzy basis functions are capable of uniformly approximating any real continuous function on a compact set to arbitrary accuracy. Note, however, that universal approximation refers to functional approximation rather than to estimation from small-sample statistics where asymptotic convergence does not hold. As a consequence, universal approximation is necessary but not sufficient, and it can become irrelevant as one attempts to learn from a limited number of observations.

According to Friedman (1994, p. 40), there are actually two ways to obtain an accurate estimate of the target function $f(x)$. One way is "to place a very restrictive set of constraints on the approximation $f^*(x)$, defining a small set (class) of eligible solutions. This approach corresponds to regularization; it is effective to the extent that a good choice of constraints becomes available and it requires knowledge (outside the data) concerning the properties of the target function." Friedman goes on to say that "in the absence of such knowledge (outside the data)—constraints—one must appeal to the second alternative for obtaining a good approximation: a training sample large enough to densely pack the input space. Although this is often feasible for input variable spaces of low dimension (few input variables), it is not possible for high-dimensional spaces, even with very large training sets. There are many manifestations of this curse of dimensionality, all of which tend to make geometrical intuition, gained in low-dimensional settings, inapplicable to higher-dimensional problems."

As an example, Friedman (1994, p. 11) lets "$n$ be the dimensionality of the input space and $N$ be the training sample size; the sampling density is then proportional to $N^{1/n}$. Thus, if $N = 100$ represents the sampling density for a single input problem, then $10^{20}$ is the sample size required to achieve the same sampling density using 10 inputs. Thus, in high dimensions, all (feasible) training samples populate the input space very sparsely. As a consequence, the interpoint distances between sample points are all large and approximately equal. Therefore, neighborhoods that contain even only a few sample points have large radii; the average (expected) edge length of a hypercubical neighborhood (volume) containing a fraction $p$ of the training points is $e_n(p) = p^{1/n}$, so that $e_{10}(.01) = 0.63$ and $e_{10}(.1) = 0.80$. The edge length corresponding to the entire training sample is 1.0. To capture 10% of the data points, one must include over 80% of the range of each input variable. Such neighborhoods are not very 'local'." One much-talked about benefit of neural networks comes from their "uncanny" ability to apparently handle the curse of dimensionality, whereas statistical methods fail on the same task. Plausible explanations for neural networks "successfully" coping with the curse have been given. Among them is the suggestion that they employ clever preprocessing (decorrelation) and clustering methods to achieve an effectively lower dimension.

Fundamental links have been recently (re)established between statistical pattern recognition and neural networks by showing the network outputs to possess specific statistical significance. As an example, backpropagation learning, the method of choice for training multilayer perceptrons (MLP) or hyperplane classifiers, is defined over the error surface $e(f) = S[y_i - f(x_i)]^2$. Geman et al. (1992) showed that, among all the functions of $x$, the regression is the best predictor of $y$ given $x$ in the minimum square error (MSE) sense, i.e., $E[(y - f(x))^2/x] \geq E[(y - E(y/x))^2/x]$. Another recent but related result, due to Richard and Lippmann (1991), states that MSE approximation estimates Bayesian probabilities according to the degree to which the training data reflect true likelihood distributions and a priori probabilities. The Geman et al. and Richards and Lippmann results were known earlier; they hold only asymptotically, when the number of training samples becomes infinite, and thus they give little insight about learning from small-sample statistics. The applicability of neural networks was also expanded to inverse problems, as they were shown to implement the equivalent of statistical maximum a posteriori (MAP) estimation.

## IV. DATA ACQUISITION

Any pattern recognition system has to sense its environment and acquire an initial representation of its immediate surroundings that is as faithful as possible. Data acquisition goes beyond raw data; e.g., a system that acquires visual data may emulate the architecture of the human (or other biological) visual system. It is not restricted to using biologically inspired sensory modalities; it may employ artificial sensing devices as well. In any case, it may derive intrinsic feature maps, including depth, lightness, and motion, or features such as sonar Doppler shift (as in bats), or possibly polarization or earth's magnetic field information for navigation purposes.

Most images are acquired using conventional visible light sensors such as TV cameras. However, in many application domains, other types of image-forming sensors are widely used. For example, in medical applications, X-ray, ultrasound, magnetic resonance, and positron emission sensing are commonly used; some of these sensing techniques can produce three-dimensional (3D), rather than two-dimensional (2D), image data. Commonly used sensors in remote sensing detect radiation in many different spectral bands, including thermal infrared; sonar and radar sensors, which yield information about range as well as reflectivity, are also widely used. Range sensors of various types are also used in short-range applications to obtain depth maps of the visible surfaces in a scene.

Some of the relevant issues in sensing and data acquisition are related to sampling and resolution. The early scale or Gaussian pyramids (Burt and Adelson, 1983; Rosenfeld, 1984) reflected the observation that pattern recognition takes place at different scales. Their immediate successors, Laplacian pyramids, combined resolution and contrast ("edge") for compression purposes. Later research has focused on joint spatial/spatial-frequency representations, including Gabor and wavelet representations (Daugman, 1983). The sampling strategies used account for the uncertainty principle by a tradeoff between spatial and spatial frequency resolution. They replicate the space-variant aspect of the human retina, where the sampling resolution decreases from the center (fovea) toward the periphery. Space-variant sampling is useful in providing a high-resolution area within a wide visual field. The retina-like charge-coupled device (CCD) sensor for active vision developed by Sandini and Tistarelli (1994) implements the space-variant sampling characteristics of foveal vision.

Conventional sensors have limited field of view and limited depth of field. Omnidirectional and omnifocus sensors (Nayar, 1997; Krishnan and Ahuja, 1996) have also been developed. (The catadioptric sensor described by Nayar has a hemispherical field of view.) Omnidirectional sensors have significant advantages for 3D motion estimation (Nelson and Aloimonos, 1988). Sensing systems that integrate information from many viewpoints are also attracting considerable interest; two examples are the CMU Dome at Carnegie-Mellon University and the Keck Laboratory for the Study of Visual Movement at the University of Maryland, where a region of space is viewed by a hemispherical array of inward-pointing cameras.

## V. FEATURE EXTRACTION

Once data acquisition has taken place, the extraction of an appropriate set of features is one of the most difficult tasks in the design of pattern recognition systems. The concept of a feature detector can be traced back to the discovery of "bug detectors" in the frog retina (Lettvin et al., 1959). As it is obvious that not all sensory information is equally probable, Barlow (1961) has advanced the concept of redundancy reduction as a design principle for sensory processing and thus for feature extraction as well. Redundancy reduction requires that the lengths of sensory messages be proportional to their information contents, given by the negative logarithms of their respective probabilities (Shannon and Weaver, 1949). One framework for redundancy reduction involves factorial codes (Barlow et al., 1989; Linsker, 1988; Atick and Redlich, 1992), in which the probability of observing a particular signal is a product of independent factors, e.g., the features that code for it (Penev and Atick, 1996). Furthermore, if the strength of the factorial code output is proportional to its information content, the code can directly represent not only the sensory signal itself, but also its likelihood. The detection of "suspicious coincidences," i.e., events or patterns, becomes much more straightforward (Barlow, 1989).

At its lowest level, the raw feature data are derived from noisy sensory data, the properties of which are complex and difficult to characterize. In addition, there is considerable interaction among low-level features that must be identified and exploited. The number of possible features, however, is so large as to prohibit systematic exploration of all but a few possible interaction types (e.g., pairwise interactions). In addition, any sort of performance-oriented evaluation of feature subsets involves building and testing the associated pattern classifier, resulting in additional overhead cost. As a consequence, a fairly standard approach is to preselect a subset of features using abstract measures believed to be relevant to characterizing important properties of good feature sets for input reconstruction, e.g., infomax, the maximization of information transmission. This (feature selection) approach has been dubbed the "filter" approach (Kohavi and Sommerfield, 1995). It is generally much less resource intensive than building and testing the associated pattern classifiers. However, it may result in suboptimal performance if the abstract features do not correlate well with actual classification performance.

Filter approaches vary considerably in how they search for good feature subsets. Because there are in general $2^N$ subsets of $N$ features, problems involving large numbers of candidate features cannot be handled by any form of systematic search. A standard technique for avoiding this combinatorial explosion simply ranks the features according to some criterion and then deletes lower-ranked features. For difficult pattern classification tasks, simple rank selection of features generally results in suboptimal classification performance because nonlinear interactions among features are ignored and criteria like orthogonality are not sufficient to guarantee good classification performance. Thus, it is difficult to develop feature space measures to guarantee optimality in classification performance. In practice, this is best achieved by employing some form of performance evaluation of the feature subsets in searching for good subsets. This approach, dubbed the "wrapper" approach, typically involves building a classifier based on the feature subset being evaluated and using the performance of this classifier as a component of the overall evaluation (Kohavi and Sommerfield, 1995). This approach should produce better classification performance than filter approaches, but it adds considerable overhead to an already expensive search process. This, in turn, usually restricts the number of alternative feature subsets one can afford to evaluate, and may also produce suboptimal results.

Focusing purely on present classification performance ignores several important issues, namely: (1) the need to minimize the number of features used for classification and limit extraction overhead; (2) the requirement that the features selected can provide a good reconstruction of the original patterns, e.g., sparse approximation; (3) the flexibility to accommodate both static and dynamic pattern landscapes; and (4) the need to guarantee some level of performance during future trials, e.g., generalization. It is difficult for a single-strategy approach to simultaneously satisfy multiple and frequently conflicting goals. These problems, and the need to exploit nonlinear interactions among features, can be addressed using more sophisticated search techniques such as genetic algorithms (GAs). They provide efficient heuristic methods for searching large spaces, as will be shown in Section IX.

Feature extraction includes both derivation and selection. Derivation involves representation, whereas selection is concerned with the choice of a meaningful subset of features for the purpose of successful pattern classification. The initial image representations

parallel the architecture of the visual cortex and include simple features (signals) and their multiscale, spectral, and/or fractal attributes. More complex features can be derived later on, including intrinsic representations such as motion, depth, lightness, and texture, as well as geometric, algebraic, and statistical invariant features. Conformal mappings, Lie transformation groups, and projective transformations, some of the means to achieve such invariance, are described by Wechsler (1990).

Feature derivation methods belong to structural, spectral or statistical, and nonparametric or connectionist methods. Mathematical morphology (Serra, 1988), characteristic of structural methods, is used for the representation of binary and gray scale images and is particularly suitable to capture the intrinsic geometry of a particular shape. Early uses of morphology include chromosome analysis. More recent applications include the extraction of features for face verification (Tefas et al., 1998) and human posture recognition (Li et al., 1998).

One of the fundamental concepts bearing on optimal signal recovery comes from information theory (Gabor, 1946). This now famous concept, known as the *uncertainty principle,* revolves around the simultaneous resolution of information in (2D) spatial and spectral terms. It is closely related to the choice of optimal lattices of receptive fields (kernel bases) for representing (decomposing) some underlying signal. As recounted by Daugman (1990), Gabor pointed out that there exists a "quantum principle" for information, which he illustrated through the construction of a spectrogram-like information diagram. The information diagram, a plane whose axes correspond to time (or space) and frequency, must necessarily be grainy, or quantized, in the sense that it is not possible for any signal or filter (and hence any carrier or detector of information) to occupy less than a certain minimal area in this plane. This minimal or quantal area reflects the inevitable tradeoff between time (space) resolution $\Delta t(\Delta s)$ and frequency resolution $\Delta\omega$ and equals the lower bound on their product. Gabor further noted that Gaussian-modulated complex exponentials offer the optimal way to encode signals (of arbitrary spectral content) or to represent them, if one wishes the code (basis functions) primitives to have both a well-defined epoch of occurrence in time (space) and a well-defined spectral identity. The code primitives are also referred to as kernels or receptive fields in analogy to biological vision. The conclusion to be drawn from the above discussion is better understood and visualized if one associates time (space) occurrence with localization, whereas frequency change is associated with the speed at which a signal changes. One cannot achieve at the same time both optimal signal localization and optimal tracking of details related to the changes taking place in spatiotemporal patterns.

The wavelet basis functions (Mallat, 1989), self-similar and spatially localized, are spatial frequency/orientation tuned kernels. They provide one possible tessellation of the conjoint spatial/spectral signal domain. The corresponding wavelet hierarchy (pyramid) is obtained as the result of orientation-tuned decompositions at a dyadic (powers of two) sequence of scales. The 2D wavelet representation $\mathbf{W}$ of the function $f(x, y)$ is then

$$W(a_x, a_y, s_x, s_y) = \int\int f(x, y)(a_x a_y)^{-1/2}\Psi^*[(x - s_x)(a_x)^{-1}, ((y - s_y)(a_y)^{-1}]$$

where $\Psi(x, y)$ is the "mother" wavelet, $a_x$ and $a_y$ are scale parameters, and $s_x$ and $s_y$ are shift parameters. The self-similar Gabor basis functions $\boldsymbol{g}$ are a special case of nonorthogonal wavelets, correspond to sinusoids modulated by a Gaussian, can be easily tuned to any bandwidth (scale) and orientation, and are defined as

$$g(x, y) = \exp\left\{-\frac{1}{2}[(x/\sigma_x)^2 + (y/\sigma_y)^2]\right\}\exp\{j2\pi[f_x x + f_y y]\}$$

where $f_x = f_0 \cos\theta$, $f_y = f_0 \sin\theta$, $f_0$ is spatial frequency, $\theta$ is the angle of spatial orientation, and the variances $\sigma_x = c/f_x$, $\sigma_y = c/f_y$ use $c$ as a scale factor. The frequency (octave) $B$ and orientation $\Omega$ (radian) half-peak bandwidths of the "daisy petal" Gabor filters are

$$B = \log[(\pi f\lambda\sigma + \alpha)/(\pi f\lambda\sigma - \alpha)]$$

$$\Omega = 2\tan^{-1}[\alpha/(\pi f\sigma)]$$

where $\lambda$ is the spatial aspect ratio, $\lambda = \sigma_x/\sigma_y$, $\sigma = \sigma_y$, $f$ is the radial center frequency, and $\alpha = [(\ln 2)/2]^{1/2}$ (Bovik et al., 1990). The self-similar Gabor wavelets are redundant due to their being nonorthogonal, leading to an overcomplete dictionary of basis functions. They are expected to provide better performance than a dictionary consisting of orthogonal bases (Daugman, 1990).

One popular technique characteristic of statistical methods and capable of deriving low-dimensional representations is principal component analysis (PCA). It has been applied, among other things, to face representation and recognition (Pentland and Choudhury, 2000). PCA is an optimal signal representation and reconstruction method that offers reduction of a large set of correlated (Gaussian) variables to a smaller number of uncorrelated components. Following its application, one derives an orthogonal projection basis that directly leads to dimensionality reduction and possibly to feature selection. Kirby and Sirovich (1990) showed that any particular face can be economically represented in the eigenface coordinate space and that any face can be approximately reconstructed by using a small set of eigenfaces and the corresponding projections (coefficients). Applying the PCA technique to face recognition, Turk and Pentland (1991) developed a well-known eigenfaces method, where the eigenfaces correspond to the eigenvectors associated with the dominant eigenvalues of the face covariance matrix.

As feature derivation encompasses both image representation and reconstruction, corresponding statistical methods have to address information content, complexity, and optimal codes. As an example, assume that one considers an ensemble of faces and that S stands for signal-to-noise ratio (SNR), the quality of reconstruction corresponding to the number $N$ of components used in the expansion of a particular face. $H$, broadly labeled as the entropy of the reconstructed image, can be computed as the average (over $N$ components) of the component (projected) squared lengths. Using a probabilistic PCA interpretation, the probability $P$ of a particular face image is $P \sim \exp\{(-\frac{1}{2})H\}$; $-\log P$, which approximates the information content of the sensory signal (Barlow, 1989), is proportional to the length of the optimal code. One can now easily see that if the entropy $H$ for some reconstructed image goes down, its likelihood goes up. One can draw $S$-$H$ diagrams and observe how much of the SNR can be obtained relatively "cheaply"; afterward, even if one increases the SNR, i.e., gets a better approximation, the reconstruction obtained is very improbable, not at all likely (Penev, 1998).

The eigenfaces define a feature space, or "face space," that drastically reduces the dimensionality of the original space. Face

identification and verification are then carried out in the reduced space. One should remember, however, that PCA is an optimal (linear) signal representation method only in the MSE sense and that the PCA-inspired features do not necessarily provide good discrimination. Nonlinear dimensionality reduction can be achieved using the self-supervised MLP architecture, where the output exemplars are forced to be identical to the input ones during training, using backpropagation learning. Self-supervised MLPs are also known as autoencoding methods, bottleneck MLP, nonlinear PCA networks (Kramer, 1991), or replicator networks (Hecht-Nielsen, 1995). Bottleneck MLP with a single hidden layer effectively performs linear PCA, even with nonlinear hidden units (Bourland and Kamp, 1988). Nonlinear PCA has been recently addressed as a kernel eigenvalue problem by Scholkopf et al. (1998).

Linear discriminant analysis (LDA), related to the Fisher linear discriminant (FLD), is yet another statistical method commonly used for pattern recognition in general, and recently also for face recognition (Swets and Weng, 1996; Etemad and Chellappa, 1997). LDA derives a projection basis that separates the different classes as far as possible and compacts the individual classes as tightly as possible. Unlike PCA, LDA differentiates between the within- and between-class scatters when deriving class-specific feature spaces. A representative LDA/FLD-based method is the Fisherfaces method of Belhumeur et al. (1997). This method specifies the face space by composing the PCA and the FLD projections, $Q = RS$, where $R$ is the PCA projection matrix and $S$ is the FLD projection matrix derived by maximizing the ratio of the between- and within-class scatters in the transformed space. The Fisherface space, defined as $Z = Q^tX$ and known as the most discriminating features (MDF) space, is superior for face recognition to the Eigenface encoding scheme, known as the most expressive features (MEF) space, only when the training images are representative of the range of face image variations; otherwise, the performance difference between MEF and MDF is not significant (Swets and Weng, 1996). The FLD procedure, when implemented in a high-dimensional PCA space, often leads to overfitting (Liu and Wechsler, 2000). Overfitting is more likely to occur in a small training sample size scenario, which is typical in face recognition (Phillips, 1999). One possible remedy for this drawback is to artificially generate additional data and thus to increase the sample size (Etemad and Chellappa, 1997).

Independent component analysis (ICA) has emerged recently as a powerful statistical solution to the problem of blind source separation (Bell and Sejnowski, 1995; Hyvarinen and Oja, 1997; Hyvarinen, 1999). It seeks a linear transformation to express a set of random variables as linear combinations of statistically independent source variables, as has been the case for factorial codes. The search criterion involves the minimization of the mutual information expressed as a function of high-order cumulants. PCA considers second-order moments only and it uncorrelates the data. ICA provides a more powerful data representation, as it accounts for higher-order statistics and distinguishes the independent source components from their linear mixtures (the observables). ICA is not restricted to second-order statistics, as was PCA; it reduces statistical dependencies and produces a sparse code useful for subsequent pattern discrimination and associative recall (Olshausen and Field, 1996). ICA seeks nonaccidental, sparse feature codes, analogous to the goal of sensory systems, "to detect redundant features and form a representation in which these redundancies are reduced and the independent features and objects are represented explicitly" (Foldiak, 1990, p. 165). The ICA of a random vector $X$ factorizes the covariance matrix, $\mathrm{Cov}(X)$, into the form $\mathrm{Cov}(X) = F\Delta F^t$, where $\Delta$ is diagonal real positive and $F$ transforms the original random vector $X$ into a new vector $Z$, where $X = FZ$. The components of the new random vector $Z$ are independent or "as independent as possible" (Comon, 1994). To derive the ICA transformation $F$, Comon developed an algorithm that consists of three operations: whitening, rotation, and normalization. ICA has been used, in the context of biometrics, for face recognition (Bartlett and Sejnowski, 1997; Liu and Wechsler, 1999) and classification of facial actions (Donato et al., 1999).

## VI. PATTERN CLASSIFIERS

Pattern classification takes over once the features have been extracted. The design of the pattern classifier includes the choice of a particular model and possibly the estimation of its probability density function. It also includes the choice of a distance or similarity metric to measure or possibly rank how close an unknown pattern is in relation to known class prototypes. Practical differences between classifiers and internal differences in how classifiers form decision regions can lead to a taxonomy consisting of probabilistic, hyperplane, kernel, and exemplar classifiers (Lippmann, 1989), or in analogy to feature extraction methods, to a taxonomy consisting of structural or inductive artificial intelligence methods, statistical pattern recognition, and connectionist methods.

Probabilistic classifiers assume a priori probability density functions (pdfs) such as Gaussians or Gaussian mixture distributions for the input features. Hyperplane classifiers derive complex decision regions using nodes that form decision boundaries in the space spanned by the inputs. MLPs, DTs, and support vector machines (SVMs) belong to this category. Kernel methods sample and approximate the input patterns using receptive fields similar to those encountered in the visual cortex and yield projection bases. Potential functions and the cerebellar model articulation controller, and more recently RBFs, sparse approximation, and (matching) projection pursuit methods are characteristic of this class of classifiers. Finally, exemplar classifiers, using unsupervised learning (such as clustering) and a predefined norm, classify unknown patterns based on the identities of their proximal and labeled neighbors. Characteristic of exemplar classifiers are methods such as $k$-nearest neighbor classifiers, memory-based reasoning, adaptive resonance theory (ART), self-organizing feature maps (SOFMs), vector quantization (VQ), and learned VQ (LVQ), which corresponds to hybrid unsupervised VQ followed by a supervised training session using labeled samples.

Examples of distance functions used include the standard Euclidean and cosine distances. Another function, the Hausdorff distance, is tolerant to perturbations. It measures proximity rather than exact superposition and it allows for flexible matching to account for small, nonrigid local distortions (Huttenlocher et al., 1993; Takacs and Wechsler, 1998). Another distance function, used in the context of pdf estimation, is the cross (relative)-entropy, or the Kullback-Leibler (KL) divergence between two pdfs. The KL divergence, also related to the mutual information between two sources, is known to be invariant to amplitude scaling and monotonic nonlinear transformations. Similarity models attempt to model psychological space and include the feature contrast model (Tversky, 1977). Instead of considering prototypes as points in a metric space, Tversky characterized them as sets of features, where similarity is measured by both the common and distinctive features of two patterns. Similarity models have been used for querying and retrieving from large image databases (Santini and Jain, 1996).

Mixture models, the method of choice for probabilistic classifiers, usually estimate the unknown pdf using EM algorithms. Nonparametric estimation, mostly based on variations of the histogram

approximation of an unknown pdf, includes Parzen windows, conceptually similar to kernel functions such as RBFs to be discussed later. By replacing the sigmoid activation function with an exponential function, probabilistic neural networks (PNNs) can compute nonlinear decision boundaries that approach the Bayes optimum (Specht, 1990). PNNs are similar in concept to both the Parzen window approach and memory-based reasoning. They can be implemented using parallel analog networks and can be incrementally built up, which is not the case for backpropagation, where learning is restarted almost from scratch when new training data become available.

One basic aim of any pattern recognition system is to construct (discrimination) rules for classifying objects, given a *training set* of objects whose class labels are known. In the formalism used in DTs, patterns are described by a fixed collection of attributes, each with its own set of discrete values. DTs are valuable tools for the description, classification, and generalization of data (Quinlan, 1993; Murthy, 1998). Several advantages of DT-based classification are pointed out by Murthy: (1) DT methods are exploratory as opposed to inferential; they are also nonparametric. As only a few assumptions are made about the model and the data distribution, DTs can model a wide range of data distributions. (2) A hierarchical decomposition implies better use of available features and computational efficiency in classification. (3) DTs perform classification by a sequence of simple, easy-to-understand tests whose semantics are intuitively clear to domain experts. The construction of DTs uses an information-theoretical approach based on entropy (Quinlan, 1993). The C4.5 algorithm suggested by Quinlan builds the DT using a top-down, divide-and-conquer approach: select an attribute, divide the training set into subsets characterized by the possible values of the attribute, and follow the same partitioning procedure recursively with each subset until no subset contains objects from more than one class. The single-class subsets then correspond to the leaves of the DT. Attribute selection and node partitioning are driven by an entropy-based *gain ratio criterion.*

Discriminant analysis, the main staple of statistical pattern classification (Fukunaga, 1972, 1990), is concerned with building two-class classifiers, assuming the data are drawn from multivariate normal probability distributions. The parameters of the densities are estimated using the maximum likelihood (ML) procedure. The resulting densities are used to construct the decision boundaries. For two known multivariate normal distributions, the optimal decision rule is a polynomial of degree two, i.e., a paraboloid. In practical problems, there are often not enough data to provide accurate estimates. One has to impose additional constraints, e.g., that the covariance matrices corresponding to the two classes are identical, which leads to a linear rather than quadratic decision rule. Cherkassky and Mulier (1998) show that in practice, when dealing with limited datasets, the linear decision rule often performs better than the quadratic decision rule, even when it is known that the two covariance matrices are not equal.

RBFs allow clustering of similar images before classifying them and thus provide the potential for developing hierarchical classifiers. The construction of an RBF network involves three different layers. The input layer consists of source nodes (sensory units). The second layer is a hidden layer whose goal is to cluster the data and reduce its dimensionality. The output layer supplies the responses of the network to the activation patterns applied to the input layer. The transformation from the input space to the hidden-unit space is nonlinear, whereas the transformation from the hidden-unit space to the output space is linear. An RBF classifier can be viewed in two ways (Ng and Lippmann, 1991). One can interpret an RBF classifier as a set of kernel functions that expand input vectors into a high-dimensional space. This approach attempts to take advantage of the mathematical fact that a classification problem cast into a high-dimensional space is more likely to be linearly separable than one in a low-dimensional space (note the similarity to SVMs). One can also interpret the RBF classifier as a function interpolation method that tries to construct hypersurfaces, one for each class, by taking a linear combination of basis functions (see also Section IX on natural bases). These hypersurfaces can be viewed as discriminant functions, where the surface has a high value for the class it represents and a low value for all others. An unknown input vector is classified as belonging to the class associated with the hypersurface with the largest output at that point. In this case, the basis functions do not serve as a basis for a high-dimensional space, but as components in a finite expansion of the desired hypersurface where the component coefficients (the weights) have to be trained.

Clustering algorithms, yet another classification method, model data distributions and assign basis function centers. They are characteristic of unsupervised connectionist learning methods such as *k*-means, SOFMs, or LVQ (Kohonen, 1988). Implicit decision boundaries are defined among pattern classes by a Voronoi partition, e.g., Dirichlet tessellation. The well-known nearest neighbor classifier corresponds to the case where the boundaries are induced by the Voronoi partition. The *k*-nearest neighbor classifier, for large *k*, is similar to the Bayes classifier. *K*-means has been shown to be a limiting case of EM optimization for a Gaussian mixture model (Bishop, 1995).

Motivated by success in speech recognition, there has been a growing interest in the use of hidden Markov models (HMMs) to model dependencies between events characteristic of human activity with the explicit goal of understanding purposeful human motion (Wren et al., 2000). HMMs encode simple events and recognize them by estimating the likelihood that the model actually produced the observation sequence. Parameterized and coupled HMMs (Oliver et al., in press) can recognize more complex events such as two mobile and interacting human subjects. HMMs require extensive and complicated training to estimate complex interactions involving several subjects. Another recent approach to the interpretation of human activity is to use Bayesian (or belief) networks (Pearl, 1988). Modeling and conditional probabilities are still needed, as is the case with HMMs, but significant simplifications can be achieved by enforcing local and contextual spatiotemporal constraints. A probabilistic framework for modeling event dependencies is provided by the Bayesian networks. Belief networks are directed acyclic graphs, usually handcrafted. The conditional probabilities associated with event dependencies connecting the nodes are learned. Recent applications of belief networks include perceptual grouping of features for human face detection (Yow and Cipolla, 1996) and human activity recognition (Intille and Bobick, 1999).

## VII. PERFORMANCE EVALUATION

Performance evaluation studies are not common because they require significant effort and cooperation among different research groups. In particular, they require the development of standard databases and common evaluation procedures to perform meaningful benchmark studies. This includes collection of large and representative databases and the design of evaluation procedures for comparing competing algorithms. The benefits of such evaluation procedures include (1) placing pattern recognition on solid experimental and scientific grounds; (2) assisting in the development of

engineering solutions to practical problems; and (3) allowing accurate assessment of the state of the art. Despite these benefits, the research community for the most part has not taken the necessary steps. There are a few exceptions; standard databases are available from the National Institute of Standards and Technology (NIST) in the areas of handwritten character recognition, and more recently, the FERET facial database. The European community (EC) under the ESPRIT program has made available the multimodal (audio and) video M2VTS (Pigeon and Vandendorpe, 1997) face database.

StatLog (Michie et al., 1994) and FERET (Phillips et al., 1998) are noteworthy examples of large benchmark studies, and workshops on evaluation techniques are starting to take place (Bowyer and Phillips, 1998). The StatLog project, sponsored by the ESPRIT (EC) initiative in the early 1990s, compared and evaluated a range of (machine learning) classification techniques and gave an assessment of their merits, disadvantages, and range of applications using comparative trials on large-scale commercial and industrial problems. The FERET evaluation procedure is an independently administered test of face recognition algorithms, where the data have been divided into development and sequestered portions. The test was designed to (1) allow a direct comparison between different algorithms; (2) identify the most promising approaches; (3) assess the state of the art in face recognition; (4) identify future directions of research; and (5) advance the state of the art in face recognition.

An interesting study by Moon and Phillips (1998) considered the comparative performance of distance functions associated with PCA-based face recognition algorithms. Their findings indicate that the best classification was achieved using Mahalanobis or combined angle and Mahalanobis. Combined L1 and Mahalanobis received the lowest score. The other distances used included L1, L2, the angle between the feature vectors, and combined L2 and Mahalanobis. More recently, Donato et al. (1999) reported on a study whose goal was to compare techniques for automatically recognizing facial actions in sequences of images. These techniques include analysis of facial motion through estimation of optical flow; holistic spatial analysis, such as PCA, ICA, local feature analysis (LFA), and LDA; and methods based on the outputs of local filters, such as Gabor wavelet representations. The best performance was obtained using the Gabor wavelet and ICA representation for classifying 12 facial actions of the upper and lower face. Donato et al. (p. 974) conclude by saying "the results provide converging evidence for the importance of using local filters, high spatial frequencies, and statistical independence for classifying facial actions."

Specific measures used for performance evaluation include distance and similarity functions (see Section VI), ranking, receiver operating characteristic (ROC) curves, confusion matrices, and $d'$ statistics (Macmillan and Creelman, 1991). The entries in confusion matrices consist of true and false positives and false and true negatives. Sensitivity (ratio of true positives to the acceptance class), specificity (ratio of true negatives to the rejection class), and accuracy (ratio of true positives and negatives to the total of both classes) are standard measures. ROC curves, plotting the false-alarm (FA) rate on the horizontal axis against the hit (true positive) rate (H) plotted vertically, are useful. They display for every value of the FA rate, ranging from 0 to 1, what the H rate would be for a particular sensitivity ($\Phi$) level. When $\Phi$ = nil, the ROC is the major diagonal (chance line) and corresponds to the case when the H and FA rates are equal. It is obvious, but not desirable, that one can increase the hit rate at will by allowing the FA rate to increase. The $d'$ statistical measure, usually used on human benchmark studies related to pattern recognition, records the discrepancy between the H and the FA rates. As an example, when subjects cannot discriminate, H = FA and $d' = 0$. Moderate pattern recognition performance implies $d' = 1$; perfect accuracy implies $d' = \infty$.

Relevant to any pattern recognition system is its ability to accept, reject, or remain undecided under the open world assumption, according to the thresholding (choice) methods employed. Adaptive thresholding, ranking, and relative confidence are several choices for making a classification decision. Another choice is to use the tools of statistical decision theory. In biometrics studies, Daugman (1993) has shown how the problem of recognizing the signature of a given iris as belonging to a particular individual, either after exhaustive search through a large database or just by comparison with a single authentication template, can be formulated within the framework of statistical decision theory. This framework also resolves the critical problem of assigning a confidence level to any such recognition decision.

Data presentation strategy during training and learning is very important. Data are usually split into training, tuning, and test sets. Tuning data are important for learning the best classifier for given training data with an eye toward improved generalization. Test data are used later to evaluate to what degree this has been achieved. Resampling techniques allow one to artificially increase the effective size of the training set and to achieve better generalization (see Section X on more recent resampling techniques characteristic of active learning, such as perturbation methods, which are performance driven). Leave-one-out, ($k$-fold) crossvalidation, and bootstrapping are examples of standard resampling techniques (Weiss and Kulikowski, 1991).

## VIII. PREDICTIVE LEARNING

One of the goals of pattern recognition is to learn classifier models whose expected performance on unseen data falls within acceptable bounds. This requirement comes from the need to predict the degree of generalization and robustness of the classifier. Generalization ability is usually based on the MSE (empirical risk) observed during training, and from making an educated guess as to the expected deviation from the empirical risk during future testing. As an example, Barron (1991) relates generalization ability to predictive learning by an expression consisting of two terms. The first component, the approximation error, refers to the distance between the target function and the closest classifier model for a given architecture. The second component, the estimation error, refers to the distance between the chosen model function and the estimated model function. It is the role of predictive learning to choose the proper structure and architecture for the classifier model, as a result of the tradeoff between overfitting, usually associated with complexity, and empirical risk (misclassification errors) on training data. For example, when the classifier is modeled using a structure consisting of polynomials, one trades between overfitting and misclassification errors by properly setting the degree of the fitting polynomials.

The goal of predictive learning is to find within such an ordered structure, based on complexity measures, the classifier model most likely to reduce the expected (predicted) risk of misclassification errors on novel data not encountered during training. This task amounts to either (1) keeping the error bounds (confidence interval) fixed and minimizing the training error (empirical risk) or (2) keeping the value of the empirical risk fixed and minimizing the confidence interval (Vapnik, 1995). Neural networks implement the first approach, whereas SVMs, characteristic of statistical learning theory (SLT), implement the second approach. Fundamental to SLT is the notion of consistency, where both the empirical risk, encoun-

tered during training, and the expected risk, for data yet unseen, converge to the minimal possible value. Probably approximately correct (PAC) learning (Valiant, 1984) is just as a particular case of the consistency concept, commonly used in statistics, in which some constraints on computational complexity were incorporated (Vapnik, 1995). If these constraints are removed from the PAC definition, one is left with nonparametric inference in the sense of statistics (Valiant, 1991) and PAC constitutes a particular case of SLT, namely, the theory of bounds (Vapnik, 1995). In particular, given $a$ and $b$, PAC will, with probability at least $(1 - a)$, produce an approximation pattern classification system $f$ such that $f$ is a $b$-approximation of the target classifier, i.e., the error $|f - f'| < b$. The bound on the error $b$ is distribution free because it must hold for any pdf of the training data.

Predictive learning requires an explicit framework to carry out its basic task, that of functional approximation. As most functional approximation problems are ill posed, additional constraints, some of them based on a priori knowledge, are needed to regularize the problem and to estimate continuous mappings from a limited number of observations. Specifically, as predictive learning is involved in inductive inference, one needs to consider a regularization framework based on inductive principles (bias), leading to better generalizations. One obvious choice comes from statistics: the Bayesian approach, where $p(u|v) = p(v|u)p(u)/p(v)$, in which $v$ and $u$ stand for observed and original data, respectively. The goal is to seek (estimate) the original data in terms of the observed data. The MAP and ML estimates seek max $p(u/v)$ and max $p(v/u)$, respectively, and for uniform $p(u)$, discarding $p(v)$, they are similar. The inductive principle underlying the Bayesian approach is that of minimum classification and risk error. The penalty (constraint) corresponding to this inductive principle consists of the a priori probability and involves information outside the training data, an obvious drawback. The Bayesian approach also suffers from its inability to fuse supporting or conflicting evidence, and to cope with missing or incomplete information.

Choosing the right model for a classifier requires an inductive principle, or bias. There are only a handful of known inductive (inference) principles, including Bayesian inference, regularization, empirical and/or structural risk minimization, and MDL (Cherkassky and Mulier, 1998; Vapnik, 1998). Bayesian inference uses additional a priori (probability) information about approximating functions in order to obtain a classifier model from the data available. This adds subjectivity to the design of a classifier, because the final model chosen depends largely on a good choice of priors. One way to estimate the unknown pdfs needed to specify the classifier, in the Bayesian framework, is to use marginalization, i.e., to average over all possible models by integrating out redundant variables, a daunting and challenging computational task (Cherkassky and Mulier, 1998). This can possibly be addressed using Monte Carlo methods (Bishop, 1995) when Gaussian assumptions do not hold. The other way to implement the Bayesian approach is to search for the MAP probability estimate. This is equivalent to the penalization formulation, characteristic of regularization methods. Choosing the value of the regularization parameter is equivalent to finding a good prior.

The main competition to the Bayesian framework comes from SLT. It takes the form of structural risk minimization (SRM), which is based on explicit minimization of the prediction risk. Under SRM, the approximating functions, ordered according to their complexity, form a nesting structure. For approximating functions linear in their parameters, their complexity is given by the number of free parameters. For nonlinear functions, their complexity is defined as the VC dimension. The referred-to structure under SRM parallels the priors structure under the Bayesian approach. SRM implements complexity control for model selection using analytic upper-bound estimates for the expected risk rather than marginalization. Most important, SRM can be applied when the true model does not belong to the set of approximating functions. The Bayesian model fails in such cases (Cherkassky and Mulier, 1998).

Complexity, the main component behind predictive learning methods, is usually associated with dimensionality. Friedman (1994) questions the wisdom of this assumption and points to the fact that univariate (sigmoid) functions can be more difficult to approximate than some original (high-dimensional) functions. In fact, one would be tempted to move to a higher-dimensional space using complex but smart nonlinear combinations of the original coordinates (features) in order to enhance separability. This is also part of the rationale behind group data handling methods (Ivakhnenko, 1971) and SVMs. Kolmogorov's theorem, introduced in Section III, answered Hilbert's 13th problem by disproving the conjecture that there are continuous functions of three variables not representable as superpositions of continuous functions of two variables. As Friedman correctly points out, what Hilbert actually conjectured, that bad (high-dimensional) functions cannot be represented in a simple way by good (low-dimensional) functions, is actually true, as the univariate functions involved in Kolmogorov's decomposition may be very wiggly and quite complex. Lorentz (1986) has also remarked that Kolmogorov's theorem shows only that the number of variables $n$ is not a satisfactory characteristic of 'badness'. The basic reason for the curse of dimensionality is that functions of high dimension can be much more complex than those of low dimension. The curse of dimensionality can be overcome for simple (intrinsically low-dimensional) functions, i.e., functions that depend (locally or globally) only on small numbers of variables (possibly after clever preprocessing). The high-dimensional but unknown function then belongs to a restricted class (of nonuniform distributions of training and testing data) for which tailor-made methods exist.

## IX. NATURAL BASES

A fairly standard approach to feature selection is to use abstract measures deemed to be relevant, such as redundancy minimization, ranging from decorrelation and minimization of the root mean square (rms) reconstruction error using PCA, to using independent features as in ICA, maximization of information transmission (infomax), and entropy. The nonaccidental properties of the world surrounding us, such as its spatiotemporal coherence, also have much to do with the design of imaging systems. Those viewpoints, formulated by Barlow (1989), also stated that adaptation and decorrelation are basic functionalities for the visual cortex. They have led to a growing interest in (a) statistical characterization of natural images (Ruderman, 1994) and (b) how the statistical properties of natural images affect the optimization of the visual system.

The possibility that the bases—kernels or receptive fields—on which raw data are projected in order to derive features could be fixed once and for all would be a major step forward, as it would eliminate the need to recompute the bases. It would also provide some consistency to the process of feature extraction and selection. As an example, encoding natural scenes has been shown to take advantage of intrinsic image statistics and to seek the derivation of a natural (universal) basis (Hancock et al., 1992; Olshausen and Field, 1996). The derived basis functions have been found to closely approximate the receptive fields of simple cells in the mammalian

primary visual cortex. Barlow (1989) has argued that such receptive fields might arise from unsupervised learning, subject to redundancy reduction or minimum entropy encoding. The receptive fields found closely resemble various derivative-of-Gaussian (DOG) functions, which are spatially localized, oriented, and bandpass-like filters. Olshausen and Field (1996) derived such receptive fields based on a criterion of sparseness. Bell and Sejnowski (1995) used an independence criterion to derive qualitatively similar results.

Sparse codes fit well with psychological and physiological evidence for parts-based representation in the brain and with computational theories of object recognition (Ullman, 1996). Lee and Seung (1999) have recently described a nonnegative matrix factorization (NMF) algorithm able to learn the parts of human faces. They have further argued that vector quantization can discover a basis consisting of prototypes, each of which is a whole face. Although the basis images for PCA are eigenfaces, some of which resemble distorted versions of the whole face, the NMF basis is radically different: its images are localized features that correspond better with intuitive notions of the parts of faces. Penev and Atick (1996) have advanced the concept of local feature analysis (LFA) as a substitute for global (holistic) PCA for face recognition. In terms of local processing for feature extraction, LFA is conceptually similar to the use of Gabor jets by Malsburg's group (Okada et al., 1998) or to the attempts to define eigenimages corresponding to specific facial landmarks such as the eyes, nose, and mouth. As the wavelet functions discussed earlier are optimally localized in both the frequency domain and the time/space domain, their use can result in a very sparse representation for a given input. Based on this observation, continuous wavelet functions were used as basis functions for a pattern recognizer taking the form of a feedforward or wavelet network (Zhang and Benveniste, 1992), possibly trained using backpropagation learning.

The search for natural bases can draw from several different but related disciplines, all of them attempting to achieve sparse functional approximation. The objective for a natural basis is that the basis should be complete and low-dimensional, and that it should allow for the efficient derivation of suitable image representations corresponding to the structure of sensory signals. Once the natural basis has been derived, no additional training is necessary and both the training images and the novel images in future tasks are represented in terms of projections along the already available natural basis. A natural basis, however, also has its drawbacks; it may be too general to properly encode for a specific task. If the patterns under consideration are human faces rather than natural images, the class of patterns to be represented is quite specific, possibly indexed by gender, ethnicity, and age. One has to learn the face space rather than a "universal" and all-encompassing natural basis. This observation also fits well with the knowledge that the "bias/variance dilemma may be circumvented if we are willing to purposely introduce bias, which then makes it possible to eliminate the variance or reduce it significantly" (Haykin, 1999, p. 29). Learning low-dimensional representations of visual patterns with extensive use of prior knowledge has also been discussed by Edelman (1999, p. 38), who claims that "Learning from examples in a high-dimensional space is computationally problematic. The problem, known as the 'curse of dimensionality', lies in the exponential dependence of the required number of examples on the number of dimensions of the representation space. Dimensionality reduction thus becomes of primary importance. The challenge, then, is to reduce dimensionality while preserving the ability of the representational system to deal with novel objects without having to come up with novel features."

On its way to deriving such a natural basis, visual system design does indeed take advantage of the structure of the surrounding environment. It is driven in particular by what Edelman (1987) calls neural Darwinism, i.e., evolution using natural selection among alternative designs. Evolution takes place by maintaining one or more populations of individuals, each of them a candidate solution competing for limited resources and rewarded according to the number of offspring in future generations. Competition is implemented via selection mechanisms that choose from the dynamically changing population resulting from the birth and death of individuals. The selection mechanisms evaluate the fitness values of individuals based on some predefined fitness function, whereas the population evolves based on interbreeding using genetic operators. When the fitness functions lack an analytical form suitable for gradient descent or the computation involved is prohibitively expensive, as is the case when the solution space is too large to search exhaustively, one alternative is to use (directed) stochastic search methods for nonlinear optimization and variable selection. The unique exploration (variations farther away from an existing population) and exploitation (minor variations on fit parents) ability of evolutionary computation guided by fitness values has made it possible to explore very complex search spaces.

As an example of such evolutionary forces at work, Reinagel and Zador (1999) report evidence that the early stages of visual processing may indeed exploit the characteristic structure of natural visual stimuli. They show that sampling of the environment is an active process and that it affects the statistics of the stimuli encountered by the fovea and by the parafovea up to eccentricities of 4 degrees. Subjects were more likely to look at image regions that had high spatial contrast. Within these regions, the intensities of nearby pixels were less correlated with each other than in images selected at random. Contrast would thus be high, and correlation would be low, whenever a sample was centered on a border between different objects. As a result, the visual system increases the entropy of its effective visual input and develops corresponding edge detectors.

One computational approach for deriving a natural basis is to use evolution, taking the form of natural selection and implemented using GAs (Mitchell, 1997a). To show how it works, we now consider the problem of learning the face space from a large and diverse population. The dimensions of the face space, to be evolved using GAs, are such that their "fitness" is driven by factors such as their classification/discrimination (cognitive) and representational (perceptual) ability, cost (number of dimensions), the (categorical) density of the resulting face space, and some measure of the tradeoff between faithful face reconstruction (representation) and the expected classification accuracy (expected risk) of the face classifier. The quality of the face space can also be driven by the diversity encountered while learning the face space. Characteristic of both coevolution and active learning methods, challenging training samples can be boosted and thus given extra weight when assessing the fitness of some possible face space.

The derivation of an optimal projection basis for face encoding has been formally addressed using evolutionary pursuit (EP; Liu and Wechsler, 1998). In analogy to (exploratory) pursuit methods from statistics, EP seeks to learn an optimal face space for the dual purpose of (1) data compression and pattern reconstruction and (2) pattern classification. The challenges that EP has successfully met on limited population types are characteristic of sparse functional approximation and SLT. Specifically, EP increases the generalization ability of the face recognizer as a result of handling the tradeoff between minimizing the empirical risk encountered during training

(performance accuracy) and narrowing the expected risk (confidence interval) in order to reduce the expected risk for unseen face images. The expected risk, corresponding to a penalty factor in regularization methods, is a measure of the generalization ability of the face classifier and corresponds to the degree of class (face) separation, e.g., the density of the face space. EP starts by projecting the original images into a lower-dimensional and whitened PCA space. Directed but random rotations of the basis vectors in this space are then searched for by GAs where evolution is driven by a fitness function defined in terms of performance accuracy (empirical risk) and class separation (confidence interval).

Learning the face space requires EP to search through a large number of possible subsets of rotated axes in a properly whitened PCA space. The rotation angles (represented by strings of bits) and the axis indicators (indicating whether or not the axes are chosen) constitute the form of the search space whose size (2 to the power of the length of the whole string) is too large to be searched exhaustively. The number and choice of (nonorthogonal) axes in the subsets and the angles of rotation are evolved using GAs. GAs work by maintaining a constant-sized population of candidate solutions known as individuals (chromosomes). The power of GAs lies in their ability to exploit, in a highly efficient manner, information about a large number of individuals. The search underlying GAs is such that breadth and depth—exploration and exploitation—are balanced according to the observed performance of the individuals evolved so far. By allocating more reproductive occurrences to above-average individuals, the overall effect is to increase the population's average fitness.

While learning the face space, evolution is driven by a fitness function formulated as follows: $\zeta(F) = \zeta_a(F) + \lambda\zeta_g(F)$, where $F$ encompasses the parameters (such as the number of axes and the angles of rotation defining each chromosome solution) subject to learning. The first term $\zeta_a(F)$ records performance accuracy, i.e., the empirical risk; the second term $\zeta_g(F)$ is the generalization index, a measure of class separation or face space density; $\lambda$ is a positive constant that indicates the importance of the second term relative to the first one. Accuracy indicates the extent to which learning has been successful so far, whereas generalization gives an indication of the expected fitness on future trials. It is interesting to note that, in analogy to SLT (Vapnik, 1995), the generalization index is conceptually similar to the capacity of the classifier and is used here to prevent overfitting. By combining those two terms with a proper weight factor $\lambda$, GAs can evolve balanced results with good recognition performance and generalization ability. The fitness function is similar to the cost functional used by regularization theory (Poggio and Girosi, 1990) and to the cost function used by sparse coding (Olshausen and Field, 1996). The cost functional of the former method exploits a regularization parameter to control the compromise between the solution's closeness to the data and the degree of regularization (quality) of the solution. The cost function of the latter method uses a positive constant to achieve a balance between an information preservation term and a term assessing the sparseness of the derived code.

## X. MIXTURES OF EXPERTS

By combining different modalities, one can enhance the performance of an identification or classification system. Modalities here are meant to include different sensors and/or classifier types. The corresponding approaches are usually referred to as data fusion and mixture of experts, respectively. As an example, by combining face, voice, and lip movement recognition, Frischholz and Dieckmann (2000) have shown how to build a highly accurate multimodal biometric identification system (BioID). In a pattern recognition context, we consider in this section the mixture of experts approach. One simple method characteristic of this approach is cross-validation, which employs a winner-take-all (WTA) combination strategy. It can be argued that WTA "wastes" those experts (models) that lose the competition. Instead of choosing a single "best" model for a given pattern recognition problem, a combination of several predictive models may produce an enhanced pattern classifier.

Model combination approaches are an attempt to capture the information made available by each of the experts. Typical model combination procedures consist of a two-stage process. In the first stage, the training data are used to separately estimate a number of different models. The parameters of these models are then held fixed. In the second stage, these models are (linearly or nonlinearly) combined, mixed, or gated to produce the final predictive model (Cherkassky and Mulier, 1998). Intuition suggests that the combination of different tests must improve performance. Daugman (2000) shows, however, that a strong biometric is better used alone than in combination with a weaker one. According to Daugman (p. 4), the key to "the apparent paradox is that when two tests are combined, one of the resulting new error rates—FA or false rejection (FR), depending on the combination rule used—becomes *better than that of the stronger of the two tests,* while the other error rate becomes *worse than even that of the weaker of the tests.* If the two biometric tests differ significantly in their power, and each operates at its own crossover point where $P(\text{FA}) = P(\text{FR})$, then combining them actually results in a significantly *worse* performance than relying solely on the one, stronger, biometric." Specific mixtures of experts architectures used for model combination usually produce a model combination by minimizing the empirical risk at each stage (Perone and Cooper, 1993) or, as is the case with stacking predictors (Wolpert, 1992), employ a resampling technique similar to cross-validation. In the first approach, the training data are first used to estimate the candidate models, and then the combined model is created by taking the weighted average. The resampling for stacking predictors is done so that the data samples used to estimate the individual approximating functions are not used to estimate the mixture coefficients.

Consider now the problem of learning a classification mapping whose form is different for different regions of the input space. Although a single homogeneous network could be applied to this problem, one expects that the task would be made easier if different expert networks are assigned to each of the different regions, while a "gating" network, which also sees the input data, decides which of the experts should determine the classification output (Bishop, 1995). Such networks, based on the "divide and conquer" modularity principle (Jacobs et al., 1991), train the expert networks and the gating networks together. The goal of the training procedure is to have the gating network learn an appropriate decomposition of the input space into different regions, and assign to individual expert networks the responsibility for making classification decisions for input vectors falling within their regions of purview. Jordan and Jacobs (1994) extend this approach by considering a hierarchical system in which each expert network can itself consist of a mixture-of-experts network, complete with its own gating network.

A similar concept using corrective training, and driven by an active learning scheme, was suggested by Krogh and Vedelsby (1995). The active learning scheme takes advantage of the obvious observation that a combination of the outputs of several networks (or other predictors) is useful only if they disagree on some inputs. The

disagreement, called the ensemble ambiguity, can then reduce the generalization error of the network ensemble. Most recently, Hinton (1999) has introduced a similar model, that of a product of experts, in order to combine multiple probabilistic models for the same data. According to Hinton, this is a very efficient way of modeling high-dimensional data that simultaneously satisfies many different low-dimensional constraints. Data vectors that satisfy one constraint but violate other constraints will be ruled out by their low probabilities under the other expert models. The overall result, that neural population codes are learned, represents yet another realization of the search for sparse and local encodings, as discussed elsewhere in this paper. Gating networks as described above can be shown to have conceptual similarities to mixture estimation and the EM algorithm (Jordan and Jacobs, 1994).

The SVM, a classification method based on structural risk minimization, is yet another manifestation of the mixture of experts approach. The input to the SVM training algorithm is a training set $(\mathbf{x}_j, y_j)$ and some kernel, possibly RBFs, acting as local experts and properly weighted (Vapnik, 1998; Scholkopf et al., 1999). The training data consists of feature vectors $\mathbf{x}_j$ and class memberships $y_j$, with the class label being either $-1$ or $+1$. SVMs seek separating hyperplanes $D(\mathbf{x})$, defined as $D(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + w_0$, by mapping the input data $\mathbf{x}$ into a higher-dimensional space $\mathbf{z}$ using a nonlinear function **g.** For an SVM, the optimal hyperplane has maximal margin; the data points at the (maximum) margin are called the support vectors because they alone define the optimal hyperplane. The reason for mapping the input into a higher-dimensional space is that this mapping leads to better class separability. The complexity of the SVM decision boundary is independent of the feature ($\mathbf{z}$) space dimensionality, which can be very large (or even infinite).

SVM optimization takes advantage of the fact that the evaluation of the inner products between the feature vectors in a high-dimensional feature space is done indirectly via the evaluation of the kernel $\boldsymbol{H}$ between support vectors and vectors in the input space, $(\mathbf{z} \cdot \mathbf{z}') = \boldsymbol{H}(\mathbf{x}, \mathbf{x}')$, where the vectors $\mathbf{z}$ and $\mathbf{z}'$ are the vectors $\mathbf{x}$ and $\mathbf{x}'$ mapped into the feature space. In the dual form, the SVM decision function has the form $D(\mathbf{x}) = \sum_{i=1}^{M} \beta_i y_i \mathbf{H}(\mathbf{x}_i, \mathbf{x}')$. The number $\boldsymbol{M}$ of RBFs, the kernel centers, which correspond to the support vectors, and the coefficients $\beta_i$ are all automatically determined by solving a quadratic optimization problem. The output of the SVM training algorithm is a set of support vectors $\mathbf{s}_i$ and weights $\alpha_i$. The support vectors are the feature vectors that characterize the boundary between the two classes. The weights are the relative contributions of the support vectors to the decision surface. The SVM decision surface $\delta$ using RBFs has the following form: $\delta(\mathbf{z}) = \sum_i \alpha_i y_i K(s_i, \mathbf{z})$, where $K$ is the RBF kernel. The weights $\alpha_i$ are the gating or mixture parameters that determine the relative influence of each support vector. SVMs have been applied to face detection (Osuna et al., 1997), eye detection (Huang et al., 1998), and face recognition (Phillips, 1999).

One can expand on the concept of a mixture of experts if, in addition to different classifier designs, one also considers different strategies or expertise for generating the training data. Learning classifiers from small training sets is difficult in that the parameters of the data distribution cannot be estimated properly. Due to the small number of training data, some of them (outliers) can greatly distort the distribution. Classifiers based on small training sets are thus usually biased or unstable (Skurichina and Duin, 1998). Bootstrap (Breiman, 1996), based on random sampling with replacement, allows one to get more accurate statistical estimates. By taking a bootstrap replicate, one is likely to avoid the outliers in the original training set. Bootstrap estimators are not always superior to leave-one-out (crossvalidation) on small samples, despite the fact that while leave-one-out is nearly unbiased, its variance can be high for small samples (Weiss and Kulikowski, 1991). Bagging, based on bootstrapping and aggregation, works by averaging the parameters of the classifiers built from several bootstrap replicates. Bagging is useful for unstable (biased and large-variance) classifiers, but it can degrade the performance of stable classifiers.

The basic paradigm for improving the accuracy of unstable methods is that of perturbing and combining. As an example, Freund and Shapire (1996) proposed an **arc**ing algorithm that **a**daptively **r**esamples and **c**ombines so that the weights during resampling are increased for those cases most often misclassified. Arcing has proved more successful than bagging in test set error reduction. Both bootstrap aggregating (bagging) and arcing (boosting) manipulate the training data in order to generate different classifiers. Combining multiple versions through either bagging or arcing then reduces the variance significantly (Breiman, 1998). An empirical comparison of voting classification algorithms has been made recently by Bauer and Kohavi (1999).

## XI. CONCLUSIONS

The ultimate objective of pattern recognition is to develop an integrated framework where feature extraction, model (classifier) selection, and predictive learning are iteratively performed with the goal of optimal (classifier) approximation. One challenge to such an effort is to derive a functional and task-oriented projection basis that does not require retraining. This basis would correspond to a compact dictionary or code book useful for efficient representation and classification. Efforts in this direction are now coming from several disciplines and are illustrated by matching pursuit methods for adaptive signal processing (Mallat and Zhang, 1993), kernel methods for neural learning (Poggio and Girosi, 1998), pursuit methods in statistics (Friedman and Stuetzle, 1981), sparse and generative models in the neurosciences (Olshausen and Field, 1996), and exploratory pursuit (Liu and Wechsler, 1998). Sparse approximation using kernel methods has recently been shown to be equivalent to SVM (Girosi, 1998) and basis pursuit (Chen et al., 1995).

Another interesting direction for further research is related to the representations stored in memory and used for pattern recognition. It is well known that caricatures of human faces are recognized better than real images. The caricaturing process is based on the selective deformation of the features of a face pattern. The caricaturing process works on images by exaggerating the most distinctive features and leaving unchanged the features that are most common. According to psychological studies, human memory stores a caricature representation (of a face) rather than the veridical image representation (Rhodes et al., 1987). Experimental work suggests that the time required for the recognition of caricatured images is less than the time required for veridical images. Properties related to memorability/distinctiveness seem to affect performance during human benchmark studies of face recognition but in a task-dependent fashion (O'Toole et al., 1993).

The memorability component of face typicality has a direct relation with the H rate and an inverse relation with the FA rate (Vokey and Read, 1995) for both face recognition memory tasks and picture recognition tasks involving faces. The task of "picture recognition," or verification, differs from face recognition in that the observer is not concerned with face identity, regardless of the image changes taking place between study and test, but is rather concerned with image identity. It appears that the primary increase in the H rate

is due to increases in encoding (gallery) distinctiveness. Increases in distinctiveness at the retrieval (probe) stage contribute to substantial reductions in the FA rate. It also has become apparent that for facial image recognition (verification), the primary locus of the caricature effect is at the retrieval stage; for face recognition (identity), its primary locus is at the encoding stage. This suggests several directions for fruitful research in pattern recognition. First, what should one caricature? How can one automate the annotation process required to provide landmarks for the distortion process involved in caricaturing? Second, and quite intriguing, is the possibility that we have to distort (caricature) both the gallery (memory) and the probe (test) representations in order to be more accurate in both identity and verification tasks.

## REFERENCES


J.J. Atick and N.A. Redlich, What does the retina know about natural scenes? Neural Computation 4 (1992), 196–210.

H.B. Barlow, "Possible principles underlying the transformation of sensory messages," Sensory Communication, W. Rosenblith (Editor), MIT Press, Cambridge, MA, 1961, pp. 217–234.

H.B. Barlow, Unsupervised learning, Neural Computation 1 (1989), 295–311.

H.B. Barlow, T.P. Kaushal, and G.J. Mitchison, Finding minimum entropy codes, Neural Computation 1 (1989), 412–423.

A.R. Barron, "Approximation and estimation bounds for artificial neural networks," Proceedings of the fourth workshop on computational learning theory, M.K. Warmuth and L.G. Valiant (Editors), Morgan Kaufmann, 1991, pp. 243–249.

M.S. Bartlett and T.J. Sejnowski, Independent component analysis of face images: A representation for face recognition, Proc fourth annual joint symp neural computation, 1997.

E. Bauer and R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, Machine Learning 36 (1999), 105–142.

J.M. Beale and F.C. Kiel, Categorical effects in the perception of faces, Cognition 57 (1995), 217–239.

P. Belhumeur, J. Hespanha, and D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, IEEE Trans Pattern Analysis Machine Intell 19 (1997), 711–720.

A.J. Bell and T.J. Sejnowski, An information maximization approach to blind separation and blind deconvolution, Neural Computation 7 (1995), 1129–1159.

C.M. Bishop, Neural networks for pattern recognition, Oxford University Press, Oxford, England, 1995.

H. Bourland and Y. Kamp, Auto-association by multilayer perceptrons and singular value decomposition, Biol Cybernet 59 (1988), 291–294.

A.C. Bovik, M. Clark, and W.S. Geisler, Multichannel texture analysis using localized spatial filters, IEEE Trans Pattern Analysis Machine Intell 12 (1990), 55–73.

K.J. Bowyer and P.J. Phillips (Editors), Empirical evaluation techniques in computer vision, IEEE Computer Society Press, Los Alamitos, CA, 1998.

L. Breiman, Bagging predictors, Machine Learning 24 (1996), 123–140.

L. Breiman, Arcing classifiers, Ann Stat 26 (1998), 801–849.

P.J. Burt and E.H. Adelson, The Laplacian pyramid as a compact image code, IEEE Trans Commun 31 (1983), 532–540.

M.A. Burton and J.R. Vokey, The face-space typicality paradox: Understanding the face-space metaphor, Q J Exp Psychol 51A (1998), 475–483.

S. Chen, D. Donoho, and M. Saunders, Atomic decomposition by basis pursuit, TR 479, Department of Statistics, Stanford University, Stanford, CA, 1995.

V. Cherkassky, J. Friedman, and H. Wechsler (Editors), From statistics to neural networks, Springer, New York, 1994.

V. Cherkassky and F. Mulier, Learning from data, Wiley, New York, 1998.

P. Comon, Independent component analysis, a new concept? Signal Processing 36 (1994), 287–314.

G. Cybenko, Approximation by superpositions of a sigmoidal function, Math Control Signals Systems 2 (1989), 303–314.

J.G. Daugman, Six formal properties of anisotropic visual filters: Structural principles and frequency/orientation selectivity, IEEE Trans Systems Man Cybernet 13 (1983), 882–887.

J.G. Daugman, "An information-theoretic view of analog representation in striate cortex," Computational neuroscience, E.L. Schwartz (Editor), MIT Press, Cambridge, MA, 1990, pp. 403–424.

J.G. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE Trans Pattern Analysis Machine Intell 15 (1993), 1148–1161.

J.G. Daugman, Biometric decision landscapes, TR 482, Computer Laboratory, Cambridge University, Cambridge, England, 2000.

G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, Classifying facial actions, IEEE Trans Pattern Analysis Machine Intell 21 (1999), 974–989.

R.O. Duda and P.E. Hart, Pattern classification and scene analysis, Wiley, New York, 1973.

R.O. Duda, P.E. Hart, and D.G. Stork, Pattern classification and scene analysis, Wiley, New York, 2000.

S. Duvdevani-Bar and S. Edelman, Visual recognition and categorization on the basis of similarities to multiple class prototypes, Int J Computer Vision 33 (1999), 201–228.

G.M. Edelman, Neural Darwinism, Basic Books, New York, 1987.

S. Edelman, Representation and recognition in vision, MIT Press, Cambridge, MA, 1999.

W.K. Estes, Concepts, categories, and psychological science, Psychol Sci 4 (1993), 143–153.

K. Etemad and R. Chellappa, Discriminant analysis for recognition of human face images, J Opt Soc Am A 14 (1997), 1724–1733.

P. Foldiak, Forming sparse representations by local anti-Hebbian learning, Biol Cybernet 4 (1990), 165–170.

Y. Freund and R.E. Shapire, "Experiments with a new boosting algorithm," Proceedings of the 13th national conference on machine learning, L. Saitta (Editor), Morgan Kaufmann, San Francisco, CA, 1996, pp. 148–156.

J.H. Friedman, "An overview of predictive learning and function approximation," From statistics to neural networks, V. Cherkassky, J. Friedman, and H. Wechsler (Editors), Springer, New York, 1994, pp. 1–61.

J.H. Friedman and W. Stuetzle, Projection pursuit regression, J Am Stat Assoc 76 (1981), 817–823.

R.W. Frischholz and U. Dieckmann, BioID: A multimodal biometric identification system, Computer 33 (Feb 2000), 64–69.

K. Fukunaga, Introduction to statistical pattern recognition, Academic Press, New York, 1972, 1990.

D. Gabor, Theory of communication, J IEE 93 (1946), 429–459.

S. Geman, E. Bienenstock, and R. Doursat, Neural networks and the bias/variance dilemma, Neural Computation 5 (1992), 1–58.

F. Girosi, An equivalence between sparse approximation and support vector machines, Neural Computation 10 (1998), 1455–1480.

S. Grossberg, Nonlinear neural networks: Principles, mechanisms, and architectures, Neural Networks 1 (1988), 17–61.

P.J.B. Hancock, R.J. Baddeley, and L.S. Smith, The principal components of natural images, Network Computation Neural Systems 3 (1992), 61–70.

S. Haykin, Neural networks, Prentice-Hall, Englewood Cliffs, NJ, 1999.

R. Hecht-Nielsen, Replicator neural networks for universal optimal source coding, Science 269 (1995), 1860–1863.

G.E. Hinton, Product of experts, Proc int conf on artificial neural networks, 1999.

J. Huang, X. Shao, and H. Wechsler, Face pose estimation using support vector machines (SVMs), Proc 14th int conf on pattern recognition, 1998, pp 154–156.

D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Trans Pattern Analysis Machine Intell 15 (1993), 850–863.

A. Hyvarinen, Survey on independent component analysis, Neural Computing Surv 2 (1999), 94–128.

A. Hyvarinen and E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation 9 (1997), 1483–1492.

S. Intille and A. Bobick, A framework for recognizing multi-agent action from visual evidence, Proc natl conf on artificial intelligence, 1999.

A.G. Ivakhnenko, Polynomial theory of complex systems, IEEE Trans Systems Man Cybernet 12 (1971), 364–378.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991), 79–87.

P. Jolicoeur, M. Gluck, and S.M. Kosslyn, Pictures and names: Making the connection, Cognitive Psychol 16 (1984), 243–275.

M.I. Jordan and R.A. Jacobs, Hierarchical mixture of experts and the EM algorithm, Neural Computation 6 (1994), 181–214.

M. Kirby and L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces, IEEE Trans Pattern Analysis Machine Intell 12 (1990), 103–108.

B.J. Knowlton and L.S. Squire, The learning of categories: Parallel brain systems for item memory and category knowledge, Science 262 (1993), 1747–1749.

R. Kohavi and D. Sommerfield, Feature subset selection using the wrapper model, Proc first int conf on knowledge discovery and data mining, 1995, pp. 192–197.

A.N. Kolmogorov, On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, Doklady, Akademii Nauk USSR 114 (1957), 679–681.

T. Kohonen, Self-organization and associative memory, Springer, New York, 1988.

M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, J AIChE 37 (1991), 233–243.

A. Krishnan and N. Ahuja, Range estimation from focus using a nonfrontal imaging camera, Int J Computer Vision 20 (1996), 169–185.

A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation and active learning," Advances in neural information processing systems (Vol. 7), D.S. Touretzky (Editor), Morgan Kaufmann, San Mateo, CA, 1995, pp. 231–238.

D.D. Lee and H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 421 (1999), 788–791.

J.Y. Lettvin, H.R. Maturana, W.S. McCulloch, and W.H. Pitts, What the frog's eye tells the frog's brain, Proc IRE 47 (1959), 1940–1959.

M.B. Lewis and R.A. Johnston, A unified account of the effects of caricaturing faces, Vis Cognition 6 (1999), 1–41.

Y. Li, S. Ma, and H. Lu, A multiscale morphological method for human posture recognition, Proc third int conf on automatic face and gesture recognition, 1998, pp. 56–61.

R. Linsker, Self-organization in a perceptual network, Computer 21 (Mar. 1988), 105–117.

R.P. Lippmann, Pattern classification using neural networks, IEEE Commun Mag (Nov. 27, 1989), 47–64.

C. Liu and H. Wechsler, Face recognition using evolutionary pursuit, Proc fifth European conf on computer vision, 1998, pp. 596–612.

C. Liu and H. Wechsler, Comparative assessment of independent component analysis (ICA) for face recognition, Proc second int conf on audio- and video-based biometric person authentication, 1999, pp. 211–216.

C. Liu and H. Wechsler, Robust coding schemes for indexing and retrieval from large face data bases, IEEE Trans Image Processing 9 (2000), 132–137.

G.G. Lorentz, Approximation of functions, Chelsea, New York, 1986.

N.A. Macmillan and C.D. Creelman, Detection theory: A user's guide, Cambridge University Press, Cambridge, England, 1991.

S. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, IEEE Trans Pattern Analysis Machine Intell 11 (1989), 674–693.

S. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans Signal Processing 41 (1993), 3397–3415.

D. Michie, D.J. Spiegelhalter, and C.C. Taylor (Editors), Machine learning, neural and statistical classification, Ellis Horwood, Chichester, England, 1994.

M. Mitchell, An introduction to genetic algorithms, MIT Press, Cambridge, MA, 1997a.

T. Mitchell, Machine learning, McGraw-Hill, New York, 1997b.

H. Moon and P.J. Phillips, "Analysis of PCA-based face recognition algorithms," Empirical evaluation techniques in computer vision, K.J. Bowyer and P.J. Phillips (Editors), IEEE Computer Society Press, Los Alamitos, CA, 1998.

S.K. Murthy, Automatic construction of decision trees from data: A multidisciplinary survey, Data Mining Knowledge Discovery 2 (1998), 345–389.

S. Nayar, Catadioptric omnidirectional camera, Proc IEEE conf on computer vision and pattern recognition, 1997, pp. 482–488.

U. Neisser, Cognition and reality, W.H. Freeman, New York, 1967.

R.C. Nelson and Y.J. Aloimonos, Finding motion parameters from spherical motion fields (or the advantages of having eyes in the back of your head), Biol Cybernet 58 (1988), 261–273.

K. Ng and R.P. Lippmann, A comparative study of the practical characteristics of neural networks and conventional pattern classifiers, TR 894, MIT Lincoln Laboratory, 1991.

K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The Bochum/USC face recognition system," Face recognition: From theory to applications, H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, and T.S. Huang (Editors), Springer, New York, 1988, pp. 186–205.

N. Oliver, B. Rosario, and A. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Trans Pattern Analysis Machine Intell (in press).

B.A. Olshausen and D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996), 607–609.

E. Osuna, R. Freund, and F. Girosi, Training support vector machines: An application to face detection, Proc IEEE conf on computer vision and pattern recognition, 1997, pp. 130–136.

A. O'Toole, H. Abdi, K.A. Deffenbacher, and D. Valentin, Low-dimensional representation of faces in higher dimensions of the face space, J Opt Soc Am A10 (1993), 405–411.

J. Pearl, Probabilistic reasoning in intelligent systems, Morgan Kaufmann, San Mateo, CA, 1988.

P.S. Penev, Local feature analysis: A statistical theory for information representation and transmission, PhD thesis, Rockefeller University, New York, 1998.

P.S. Penev and J.J. Atick, Local feature analysis: A general statistical theory for object representation, Network Computation Neural Systems 7 (1996), 477–500.

A. Pentland and T. Choudhury, Face recognition for smart environments, Computer 33 (Feb. 2000), 50–55.

M.P. Perone and L.N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," Artificial neural networks for speech and vision, R.J. Mammone (Editor), Chapman and Hall, New York, 1993, pp. 126–142.

P.J. Phillips, "Support vector machines applied to face recognition," Advances in neural information processing system (Vol. 11), M.S. Kearns, S.A. Solla, and D.A. Cohn (Editors), MIT Press, Cambridge, MA, 1999, pp. 803–809.

P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, The FERET database and evaluation procedure for face recognition algorithms, Image Vision Computing 16 (1998), 295–306.

S. Pigeon and L. Vandendorpe, The M2VTS multimodal face database (release 1.00), Proc first int conf on audio- and video-based biometric person authentication, 1997, pp 403–409.

T. Poggio, A theory of how the brain might work, Proc Cold Spring Harbor symposia on quantitative biology LV, 1990, pp. 899–910.

T. Poggio and F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks, Science 247 (1990), 978–982.

T. Poggio and F. Girosi, A sparse representation for functional approximation, Neural Computation 10 (1998), 1445–1454.

R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, San Mateo, CA, 1993.

P. Reinagel and A.M. Zador, Natural scene statistics at the centre of gaze, Network Computation Neural Systems 10 (1999), 341–350.

G. Rhodes, S. Brennan, and S. Carey, Identification and ratings of caricatures: Implications for mental representations of faces, Cognitive Psychol 19 (1987), 437–497.

G. Rhodes, S. Carey, G. Byatt, and F. Proffitt, Coding spatial variations in faces and simple shapes: A test of two models, Vision Res 38 (1998), 2307–2321.

M.D. Richard and R.P. Lippmann, Neural network classifiers estimate Bayesian a posteriori probabilities, Neural Computation 3 (1991), 461–483.

B. Ripley, Pattern recognition and neural networks, Cambridge University Press, Cambridge, England, 1996.

A. Rosenfeld (Editor), Multiresolution image processing and analysis, Springer, New York, 1984.

D.L. Ruderman, The statistics of natural images, Network Computation Neural Systems 5 (1994), 517–548.

G. Sandini and M. Tistarelli, Recognition by using an active/space-variant sensor, Proc IEEE conf on computer vision and pattern recognition, 1994, pp. 833–837.

S. Santini and R. Jain, The graphical specification of similarity queries, J Vis Lang Computing 7 (1996), 403–421.

B. Scholkopf, C.J.C. Burges, and A.J. Smola (Editors), Advances in kernel methods: Support vector machines, MIT Press, Cambridge, MA, 1999.

B. Scholkopf, A. Smola, and K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998), 1299–1319.

J. Serra, Image analysis and mathematical morphology, Academic Press, London, 1988.

C.E. Shannon and W. Weaver, The mathematical theory of communication, University of Illinois Press, Chicago, 1949.

R.N. Shepard, Cognitive psychology: A review of the book by U. Neisser, Am J Psychol 81 (1968), 285–289.

M. Skurichina and R.P.W. Duin, Bagging for linear classifiers, Pattern Recognition 31 (1998), 909–930.

D.F. Specht, Probabilistic neural networks, Neural Networks 3 (1990), 109–118.

D.L. Swets and J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans Pattern Analysis Machine Intell 18 (1996), 831–836.

B. Takacs and H. Wechsler, Fast searching of digital face libraries using binary image metrics, Proc 14th int conf on pattern recognition, 1998, pp. 1235–1237.

A. Tefas, C. Kotropoulos, and I. Pitas, Face verification based on morphological shape decomposition, Proc third int conf on automatic face and gesture rcognition, 1998, pp. 36–41.

M. Turk and A. Pentland, Eigenfaces for recognition, J Cognitive Neurosci 3 (1991), 71–86.

A. Tversky, Features of similarity, Psychol Rev 84 (1977), 327–352.

S. Ullman, High-level vision, MIT Press, Cambridge, MA, 1996.

T. Valentine, A unified account of the effects of distinctiveness, inversion, and race in face recognition, Q J Exp Psychol 43A (1991), 161–204.

L.G. Valiant, A theory of learnability, Communications ACM 27 (1984), 1134–1142.

L.G. Valiant, A view of computational learning theory, Computation and Cognition, 36 (1991).

V. Vapnik, The nature of statistical learning theory, Springer, New York, 1995.

V. Vapnik, Statistical learning theory, Wiley, New York, 1998.

J.R. Vokey and J.D. Read, "Memorability, familiarity, and categorical structure in the recognition of faces," Cognitive and computational aspects of face recognition, T. Valentine (Editor), Routledge, New York, 1995.

L.X. Wang and J.M. Mendel, Fuzzy basis functions, universal approximation, and orthogonal least-squares learning, IEEE Trans Neural Networks 3 (1992), 807–814.

H. Wechsler, Computational vision, Academic Press, New York, 1990.

S.M. Weiss and C.A. Kulikowski, Computer systems that learn, Morgan Kaufmann, San Mateo, CA, 1991.

D. Wolpert, Stacked generalization, Neural Networks 5 (1992), 241–259.

C.R. Wren, B.P. Clarkson, and A. Pentland, Understanding purposeful human motion, Proc fourth int conf on automatic face and gesture recognition, 2000, pp. 378–383.

K.C. Yow and R. Cipolla, A probabilistic framework for perceptual grouping of features for human face detection, Proc second int conf on automatic face and gesture recognition, 1996, pp. 16–21.

Q. Zhang and A. Benveniste, Wavelet networks, IEEE Trans Neural Networks 3 (1992), 889–898.