LETTER Communicated by Heinrich Buelthoff

Receptive Field Structures for Recognition

Benjamin J. Balas
bjbalas@mit.edu
Pawan Sinha
psinha@mit.edu
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02142, U.S.A.

Localized operators, like Gabor wavelets and difference-of-gaussian filters, are considered useful tools for image representation. This is due to their ability to form a sparse code that can serve as a basis set for high-fidelity reconstruction of natural images. However, for many visual tasks, the more appropriate criterion of representational efficacy is recognition rather than reconstruction. It is unclear whether simple local features provide the stability necessary to subserve robust recognition of complex objects. In this article, we search the space of two-lobed differential operators for those that constitute a good representational code under recognition and discrimination criteria. We find that a novel operator, which we call the *dissociated dipole*, displays useful properties in this regard. We describe simple computational experiments to assess the merits of such dipoles relative to the more traditional local operators. The results suggest that nonlocal operators constitute a vocabulary that is stable across a range of image transformations.

1 Introduction

Information theory has become a valuable tool for understanding the functional significance of neural response properties. In particular, the idea that a goal of early sensory processing may be to efficiently encode natural stimuli has generated a large body of work describing the function of the human visual system in terms of redundancy reduction and maximum-entropy responses (Attneave, 1954; Barlow, 1961; Atick, 1992; Field, 1994).

In the compound eye of the fly, for example, the contrast response function of a particular class of interneuron approximates the distribution of contrast levels found in natural scenes (Laughlin, 1981). This is the most efficient encoding of contrast fluctuations, meaning that from the point of view of information theory, these cells are optimally tuned to the statistics of their environment. In the context of the primate visual system, it has been proposed that the receptive fields of various cells may have the form they do for similar reasons. Olshausen and Field (1996, 1997) and Bell and Sejnowski (1997) have demonstrated that the oriented edge-finding receptive fields that are found in early visual cortex (Hubel & Wiesel, 1959) may exist because they provide an encoding of natural scenes that maximizes information. Olshausen and Field were able to produce such filters through enforcing sparseness constraints on their encoding while ensuring that the representation allowed high-fidelity reconstruction of the original scene. Bell and Sejnowski enforced the statistical independence of the filters rather than working with an explicit sparseness criterion. These two approaches are actually equivalent, as demonstrated by Olshausen and Field. An aspect of Bell and Sejnowski's work that sets it apart, however, is their progression through constraints of different strength, such as principal component analysis (orthogonal basis), ZCA (zero-phase whitening filters), and finally independent component analysis (statistical independence). These different constraints lead to qualitatively different filters, such as checkerboard-like structures and center-surround functions, resembling the preferred stimuli of cells found in some parts of the visual pathway (V4 and the lateral geniculate nucleus, (LGN), respectively).

The search for efficient codes has helped direct the efforts of researchers interested in explaining neural response properties in the visual system and fostered the study of ecological constraints in natural scenes (Simoncelli & Olshausen, 2001). However, there are many other tasks that the visual system must accomplish, for which the goal may be quite different from high-fidelity input reconstruction. The task of recognizing complex objects is an important case in point. A priori, we cannot assume that the same computations that result in sparse coding would also support robust recognition. Indeed, the resilience of human recognition performance to image degradations suggests that image measurements underlying recognition can survive significant reductions in reconstruction quality. Extracting measurements that are stable against ecologically relevant transformations of an object (lighting and pose, for example) is a constraint that might result in qualitatively different receptive field structures from the ones that support high-fidelity reconstruction.

In this article, we examine the nature of receptive fields that emerge under a recognition- rather than reconstruction-based criterion. We develop and illustrate our ideas primarily in the context of human faces, although we expect that similar analyses can be conducted with other object classes as well. In this analysis, we note the emergence of a novel receptive field structure that we call the *dissociated dipole*. These dipoles (or "sticks") perform simple nonlocal luminance comparisons, allowing a region-based representation of image structure.

We also compare the stability characteristics of various kinds of filters. These include model neurons with receptive field structures like those found by sparse coding constraints and sticks operators. Our goal is to eventually gain an understanding of how object representations that are useful for recognition might be constructed from simple image measurements.

2 Experiment 1: Searching for Simple Features in the Domain of Faces

We begin by investigating what kinds of simple features can be used to discriminate among frontally viewed faces. The choice of a specific example class is primarily for ease of exposition. The ideas we develop are intended to be more generally applicable. (We substantiate this claim in experiment 2 when we describe computational experiments with arbitrary object classes.)

Computationally, there are many methods for performing the face discrimination task with relatively high accuracy, especially if the faces are already well normalized for position, pose, and scale. Using nothing more than the Euclidean distance between faces to do nearest-neighbor classification in pixel space, one can obtain reasonably good results (∼65% with a 40-person classification task using the ORL database, compiled by AT&T Laboratories, Cambridge, UK). Using eigenfaces, one can improve this score somewhat by removing the contribution of higher-order eigenvectors, effectively "denoising" the face space. Further adjustments can be made as well, including the explicit modeling of intra- and interpersonal differences (Moghaddam, Jebara, & Pentland, 2000) and the use of more complex classifiers. On the other side of the spectrum from these global techniques are methods for rating facial similarity that rely on Gabor jets placed at fiducial points on a face (Wiskott, Fellous, Kruger, & von der Malsburg, 1997). These techniques use information at multiple spatial scales to produce a representation built up from local analyses; they are also quite successful.

The overall performance of these systems depends on both the choice of representation and the back-end classification strategy. Since we focus exclusively on the former, our goal is not to produce a system for recognition that is superior to these approaches, but rather to explore the space of front-end feature choices. In other words, we look within a specific set of image measurements, bilobed differential operators, to see what spatial analyses lead to the best invariance across images of the same person. For our purposes, a *bilobed differential operator* is a feature type in which weighted luminance is first calculated over two image regions, and the final output of the operator is the signed difference between those two average values. In general, these two image regions need not be connected. Some examples of these filters are shown in Figure 1.

Conceptually, the design of our experiment is as follows. We exhaustively consider all possible bilobed differential operators (with the individual lobes modeled as rectangles for simplicity). We evaluate the discrimination performance of the corresponding measurements over a face database (discriminability refers to maximizing separation between individuals and minimizing distances within instances of the same person). By sorting the

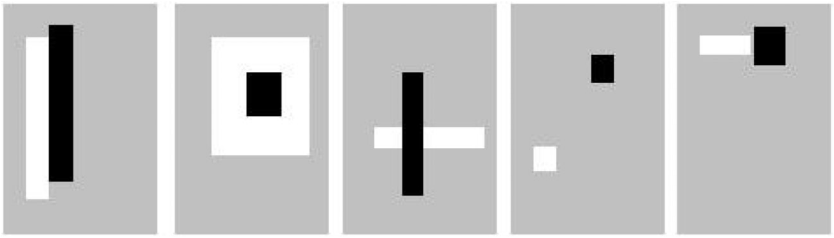

Figure 1: Examples of bilobed differential operators of the sort employed in experiment 1.

large space of all operators using the criterion of discriminability, we can determine which are likely to constitute a good vocabulary for recognition.

We note that this approach differs substantially from efforts to find reliable features for face and object detection in cluttered backgrounds. For example, Ullman's work on features of intermediate complexity (IC) (Ullman, Vidal-Naquet, & Sali, 2002) demonstrates a method for learning class-diagnostic image fragments using mutual information. These IC features are both very likely to be present in an image when the object is present and unlikely to appear in the image background by chance. Other feature learning studies have concentrated on developing generative models for object recognition (Fei-Fei, Fergus, & Perona, 2003; Fergus, Perona, & Zisserman, 2003; Fei-Fei, Fergus, & Perona, 2004) in which various appearance densities are estimated for diagnostic image fragments. This allows recognition of an object in a cluttered scene to proceed in a Bayesian manner.

These studies are unquestionably valuable to our understanding of object recognition. Our goals in this study are slightly different, however. First, we are interested in discovering what features support invariance to a particular object rather than a particular object class. It is for this reason that we do not attempt to segment the objects under consideration from a cluttered background. We envision segmentation proceeding via parts-based representations such as those described above. Indeed, it has recently been shown that simple contrast relationships can be used to detect objects in cluttered backgrounds with good accuracy (Sinha, 2002) and that good segmentation results can be obtained once one has recognized an object at the class level (Borenstein & Ullman, 2002, 2004). While it may be possible to learn diagnostic features of an individual that could be used for segmentation purposes, we believe that it is also plausible to consider segmentation as a process that proceeds prior to individuation (subordinate-level classification). Second, rather than looking for complex object parts that support invariance, we commence by considering very simple features. This means that we are not likely to find globally optimal features for individuation. Instead, we aim to determine what structural properties of potentially low-level RFs contribute to recognition. In a sense, we are trying to understand what computations between the lowest and highest levels of visual processing lead to the impressive invariances for object transformations displayed by our visual system.

Given that we are attempting to understand how recognition abilities are built up from low-level features, one might ask why we do not explicitly assume preprocessing by center-surround or wavelet filters. Indeed, others have pursued this line of thought (Edelman, 1993; Schneiderman & Kanade, 1998; Riesenhuber & Poggio, 1999), and such an analysis could help us understand how the outputs of early visual areas (such as the LGN and V1) serve as the basis for further computations that might support recognition. That said, we have chosen not to adopt this strategy, so that we can remain completely agnostic as to what basic computations are necessary first steps toward solving high-level problems. However, it is straightforward to extend this work to incorporate a front-end comprising simple filters.

2.1 Stimuli. We use faces drawn from the ORL database (Samaria and Harter, 1994) for this initial experiment. The images are all 112 × 92 pixels in size, and there are 10 unique images of each of the 40 individuals included in the database. We chose to work with 21 randomly chosen individuals in the database, using the first 5 images of each person. The faces are imaged against uniform backdrops. Therefore, the task in our experiment is not to segregate faces from a cluttered background, but rather to individuate them.

2.2 Preprocessing

2.2.1 Block Averaging. Relaxing locality constraints results in a very large number of allowable square differential operators in a particular image. To reduce the size of our search space, we first down-sample all of the images in our database to a much smaller size of 11 × 9 pixels. Much of the information necessary for successful classification is present at this small size, as evidenced by the fact that the recognition performance of a simple nearest-neighbor classifier actually increases slightly (from 65% correct at full resolution to 70% using 8 × 8 pixel blocks) if we use these smaller images as input.

2.2.2 Constructing Difference Vectors. Our next step involves changing our recognition problem from a 21-class categorization task into a binary one. We do this by constructing difference vectors, which comprise two classes of intra- and interpersonal variation (Moghaddam et al., 2000). Briefly, we subtract one image from another, and if the two images used depicted the same individual, then that difference vector captures intrapersonal variation. If the two images were of different individuals, then that difference vector would be one that captured interpersonal variation. Given these two sets, we look for spatial features that can distinguish between these two types of variation in facial appearance rather than attempting to find features that are always stable within each of 21 categories. To assemble the difference vectors used in this experiment, we took all unique pair-wise differences between images that depicted the same person (intrapersonal set) and used the first image of each individual to construct a set of pair-wise differences that matched our first set in size (interpersonal set). The faces used to construct these difference vectors were not precisely registered. We attempted to find features robust to the variations in facial position and view that arise in this data set.

2.2.3 Constructing Integral Images. Now that we have two sets of low-resolution difference vectors, we introduce one last preprocessing step designed to speed up the execution of our search. Since the differential operators we are analyzing have rectangular lobes, we construct integral images (Viola & Jones, 2001) from each of our difference vectors. Integral images allow the fast computation of rectangular image features, reducing the process to a series of look-ups. The value of each pixel in the integral image created from a given stimulus represents the sum of all pixels above and to the left of that pixel in the original picture.

2.3 Feature Ranking. In our 11 × 9 images, there are a total (n) of 2970 unique box features. Given that we are interested in all possible differential operators, there are approximately 4.5 million spatial features ($n^2/2$) for us to consider. To decide which of these features were best for recognition, we used A' as our measure of discriminability (Green & Swets, 1966). A' is a nonparametric measure of discriminability calculated by finding the area underneath an observer's ROC (receiver-operating-characteristic) curve. This curve is determined by plotting the number of "hits" and "false alarms" a given observer obtains when using a particular numerical threshold to judge the presence or absence of a signal.

In this experiment, we treat each differential operator as one observer. The signals we wish to detect are the intrapersonal difference vectors. The response of each operator (mean value of pixels under the white rectangle minus mean value of pixels under the black rectangle) was calculated on each difference vector, and then the labels associated with those vectors (intra- versus interpersonal variation) were sorted according to that numerical output. With the distribution of labeled difference vectors in hand for a particular feature, we could proceed to calculate the value of A'. We determined how many hits and false alarms there would be for a threshold placed at each possible location along the continuum of observed feature values. This allowed us to plot a discretized ROC curve for each feature. Calculating the area underneath this curve is straightforward, yielding the discriminability for that operator. A' scores range from 0.5 to 1. A perfect separation of intra- and interpersonal difference vectors would lead to an

A' score of 1, while a complete enmeshing of the two classes would lead to a score of 0.5.

In one simulation, the absolute value of each feature was taken (rectified results), and in another the original responses were unaltered (unrectified results). In this way, we could establish how instances of each class were distributed with respect to each spatial feature, both with and without information concerning the direction of brightness differences.

It is important to note at this stage that there is no reason to expect that any of the values we recover from our analysis of these spatial features will be particularly high. In boosting procedures, it is customary to use a cascade of relatively poor filters to construct a classifier capable of robust performance, meaning that even with a collection of bad features, one can obtain worthwhile results. In this experiment, we are interested only in the relative ranking of features, though it is possible that the set of features we obtain could be useful for recognition despite their poor abilities in isolation. We shall explicitly consider the utility of the features discovered here in a recognition paradigm presented in experiment 2.

2.4 Results

2.4.1 Differential Operators. The top-ranked differential operators recovered from our analysis of the space of possible two-lobed box filters are displayed in Figure 2. As we expected, the A' measured for each individual feature is not particularly high, with the best operator in these two sets scoring approximately 0.71.

There are four main classes of features that dominate the top 100 differential operators. First, features resembling center-surround structures appear in several top slots, in both the rectified and unrectified data. This is somewhat surprising, given that cells with this structure are most commonly associated with very early visual processing implicated in low-level tasks such as contrast enhancement, rather than higher-level tasks like recognition. Of course, the features we have recovered here are far larger in terms of their receptive field than typical center-surround filters used for early image processing, so perhaps these structures are useful for recognition if scaled up to larger sizes.

The second type of feature that is very prevalent in the results is what we will call a dissociated dipole, or stick, operator, and appears primarily in the unrectified results. These features have a spatially disjoint structure, meaning that they execute brightness comparisons across widely separate parts of an image. Admittedly, the connection between these operators and the known physiology of the primate visual system is weak. To date, there have been no cells with this sort of dissociated receptive field structure found in the human visual pathway, although they may exist in the auditory and somatosensory processing streams (Young, 1984; Chapin, 1986).

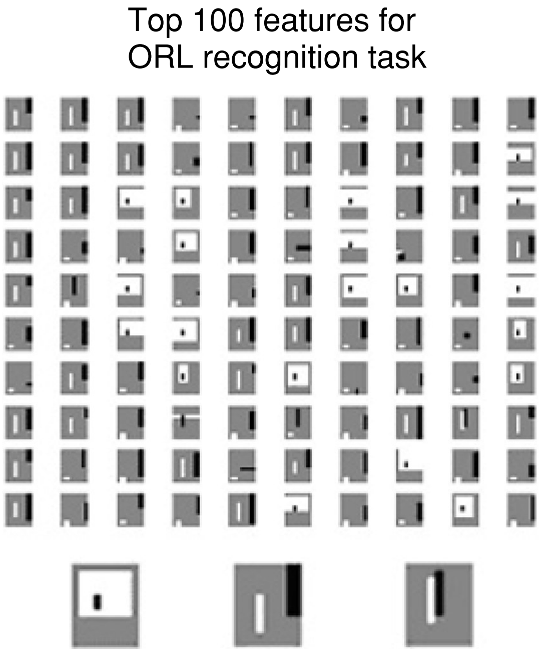

Figure 2: The top 100 ranked features for discriminating between intra- and interpersonal difference vectors. Beneath the 10 × 10 array are representatives of the most common features discovered.

The final two features are elongated edge and line detectors, which dominate the results of the rectified operators. An elongated edge detector appears in the unrectified rankings as well, but other structurally similar features are found only in the next 100 ranked features. These structures resemble some of the receptive fields known to exist in striate cortex, as well as the wavelet-like operators that support sparse coding of natural scenes.

We point out that multiple copies of these features appear throughout our rankings, which is to be expected. Small structural changes to these filters only slightly alter their A' score, meaning that many of the top features have very similar forms. We do not attribute any particular importance to the fact that the nonlocal operators that perform best appear to be comparing values on the right edge of the image to values in the center, or to the tendency for elongated edge detectors to appear in the center of the image. It is only the generic structure of each operator that is important to us here.

2.4.2 Single Rectangle Features. We chose to examine differential operators in our initial analysis for several reasons. First, cells with both excitatory and inhibitory regions are found throughout the visual system. Second, by

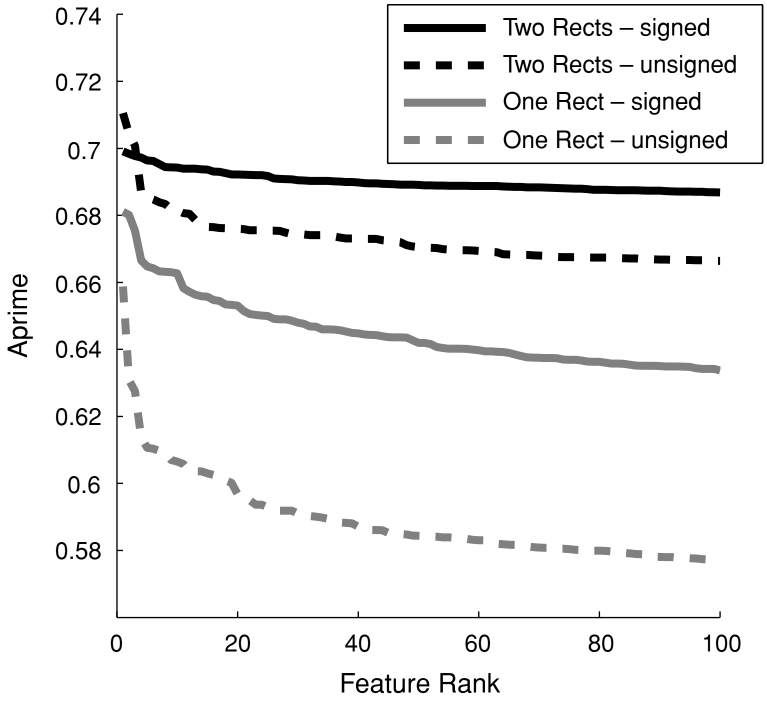

Figure 3: Plots of A' scores across the best features from each family of operators (single versus double rectangle features, as well as rectified versus unrectified operator values).

taking the difference in luminance between one region or another, one is far less sensitive to uniform changes in illumination brought on by haze a bright lighting, for example. However, given that we are using a database of faces that is already relatively well controlled in terms of lighting and pose, it may be the case that even simpler features can support recognition. To examine this possibility, we conduct the same analysis described above for differential operators on the set of all single-rectangle box features in our images.

We find that single-rectangle features are not as useful for discriminating between our two classes as are differential operators. The range of A' values for the top 100 features from each category is plotted in Figure 3, where it is clear that both sets of differential operators provide better recognition performance than single box filters. Even in circumstances where many of the reasons to employ differential operators have been removed through clever database construction (say, by disallowing fluctuations in ambient illumination), we find that they still outperform simpler measurements.

2.5 Discussion. In our analysis of the best differential operators for face recognition, we have observed a new type of operator (the dissociated dipole) that offers an alternative form of processing by which within-class stability might be achieved for images of faces. An important question to consider is how this operator fits within the framework of previous computational models of recognition, as well as whether it has any relevance to human vision.

The dissociated dipole is an instance of a higher-order image statistic, a binary measurement. The notion that such statistics might be useful for pattern recognition is not new; indeed, Julesz (1975) suggested that needle statistics could be useful for characterizing random dot textures. In the computer vision community, nonlocal comparisons are employed in integral geometry to characterize shapes (Novikoff, 1962). The possibility that nonlocal luminance comparisons may be useful for object and face recognition has not been thoroughly explored, however. Such an approach differs from traditional shape-based approaches to object recognition, in that it implicitly considers relationships between regions to be of paramount importance. Our recent results (Balas & Sinha, 2003) have demonstrated that such a nonlocal representation of faces provides for better recognition performance than a strictly local one. Furthermore, Kouh and Riesenhuber (2003) have found that to model the responses of V4 neurons to various gratings using a hierarchical model of recognition (Riesenhuber & Poggio, 1999), it is necessary to pool responses from spatially disjoint low-level neurons.

Before proceeding, we wish to specify more precisely the relationship between local, nonlocal, and global image analysis. We consider local analyses those in which a contiguous set of pixels (either 4- or 8-connected) is represented in terms of a single output value. A global analysis is similar to this, save for the amount of the image under consideration. In the limit, a global image analysis uses all pixels in the image to construct the output value. A local analysis might use only some small percentage of image area. This distinction is not truly categorical. Rather, there is a spectrum between local and global image analysis.

Likewise, a similar spectrum exists between local and nonlocal analysis. While a local analysis considers only a set of contiguous pixels, a nonlocal analysis breaks this condition of contiguity. In the extreme, one can imagine a highly nonlocal feature composed of two pixels located at opposite corners of an image. At the other extreme would be a highly local feature consisting of two neighboring pixels. Of course, there are many operators spanning these two possibilities that are neither purely local nor nonlocal. Moreover, if one measures local features (like Gabor filter outputs) at several nonoverlapping positions, is this a local or a nonlocal analysis? If one is merely concatenating the values of each local analysis into one feature vector, then this is not a truly nonlocal computation by our definition. If, however, the values of those local features are explicitly combined to

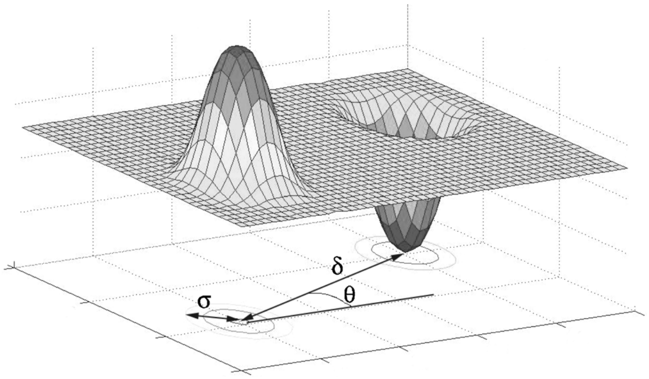

Figure 4: A dipole measurement is parameterized in terms of the space constant σ of each lobe, the distance δ between the centers of each lobe, and the angle of orientation, θ.

produce one output value, then we would have arrived at a nonlocal analysis of the image. Nonlocal analysis of this type has traditionally received less attention than local or global strategies of image processing.

The reason nonlocal representations of brightness have not been studied in great detail may be due to the sheer number of generic binary statistics. In general, the trouble with appeals to higher-order statistics for recognition is that there is a vast space of possible measurements that are allowable with the introduction of new parameters (in our case, the distance between operator lobes). This combinatorial explosion makes it hard to determine which particular measurements are actually useful within the large range of possibilities. This is, of course, a serious problem in that the utility of any set of proposed measurements is dependent on the ability to separate helpful features from useless ones.

We also note that there are several computational oddities associated with nonlocal operators. Suppose that we formulate a dissociated dipole as a difference-of-offset-gaussians operator (a model we present in full in the next experiment), allowing the distance between the two gaussians to be manipulated independent of either one's spatial constant (see Figure 4). In so doing, we lose the ability to create steerable filters (Freeman & Adelson, 1991), meaning that to obtain dipoles at a range of orientations, we have no other option than to use a large number of operators. This is not impossible, but it lacks the elegance and efficiency of more traditional approaches by which multiscale representations can be created at any orientation through the use of a small number of basis functions.

Another important difference between local and nonlocal computations is the distribution of operator outputs. Natural images are spatially redundant, meaning that the output of most local operators is near zero (Kersten, 1987). The result is a highly kurtotic distribution of filter outputs, indicating that a sparse representation of the image using those filters is expected. In many cases, this is highly desirable from both metabolic and computational viewpoints. As we increase the distance between the offset gaussians we use to model dissociated dipoles, the kurtosis of the distribution decreases significantly. This means that using these operators yields a coarse (or distributed) encoding of the image under consideration. This may not be unreasonable, especially given that distributed representations of complex objects may help increase robustness to image degradation. However, it is important to note that nonlocal computations depart from some conventional ideas about image representation in significant ways.

Finally, given that we have discussed our findings in the context of discovering receptive field structures that are good for recognition rather than encoding, it is important to describe what differences we see between those two processes. The initial stages of any visual system have to perform transduction—transforming the input into a format amenable to further processing. *Encoding* is the process by which this re-representation of the visual input is accomplished. *Recognition* is the process by which labels that reflect aspects of image content are assigned to images. The constraints on encoding processes are twofold: the input signal should be represented both accurately and efficiently. Given the variety of visual tasks that must be accomplished with the same initial input, it makes sense that early visual stages would not be committed to optimizing any one of them. For that reason, we suggest that recognition operates on a signal that is initially encoded via localized edge-like operators, but may rely on different measurements extracted from that signal that prove more useful.

In our next experiment, we directly address the question of whether the structures we have discovered in this analysis are useful for face and object classification. In this next analysis, we remove many of the simplifications necessary for an exhaustive search to be tractable in experiment 1. We also move beyond the domain of face recognition to include multiple object classes in our recognition task.

3 Experiment 2: Face and Object Recognition Using Local and Nonlocal Features

In our first experiment, we noted the emergence of center-surround operators and nonlocal operators under a recognition criterion for frontally viewed faces. However, in our first experiment, many compromises were made in order to conduct an exhaustive search through the space of possible operators. First, our images were reduced to an extremely small size in order to limit the number of features we needed to consider. Though faces can be recognized at very low resolutions, it is also clear that there is interesting and useful structure at finer spatial scales. Second, we chose to work with difference images rather than the original faces. This allowed

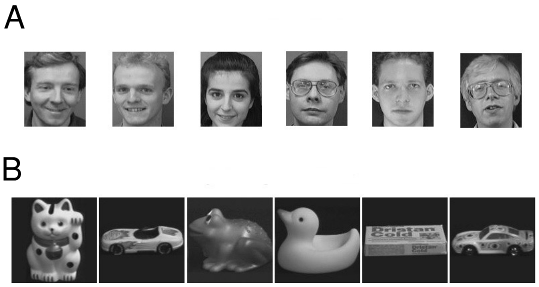

Figure 5: Examples of stimuli used in experiment 2. (A) Training images of several individuals depicted in the ORL database. (B) Training images of objects depicted in the COIL database. Note that the COIL database contains multiple exemplars of some object classes (such as the cars in this figure), making within-class discrimination a necessary part of performing recognition well using this database.

us to transform a multicategory classification task into a binary task, but embodied the implicit assumption that a differencing operation occurs as part of the recognition process. Third, we point out that in any consideration of all possible bilobed features in an image, the number of nonlocal features will far exceed the number of local features. Greater numbers need not imply better performance, yet it is still possible that the abundance of useful nonlocal operators may be a function of set size. Finally, we note that in considering only face images, it is unclear whether the features we discovered are useful for general recognition purposes or specific to face matching.

In this second experiment, we attempt to address these concerns through a recognition task that eliminates many of these difficulties. We employ high-resolution images of both faces and various complex objects in a classification task designed to test the efficacy of center-surround, local-oriented, and nonlocal features in an unbiased fashion.

3.1 Stimuli. For our face recognition experiment, we once again make use of the ORL database. In this case, all 40 individuals were used, with one image of each person serving as a training image. The images were not preprocessed in any way and remained at full resolution (112 × 92 pixels).

To help determine if our findings hold up across a range of object categories, we also conduct this recognition experiment with images taken from the COIL database (see Figure 5; Nayar, Nene, & Murase, 1996; Nene, Nayar, & Murase, 1996). These images are 128 × 128 pixel images of 100 different objects, including toy cars, foods, pharmaceutical products, and many other diverse items. We selected these images for the wide range of surface and structural properties represented by the objects. Also, repeated exemplars of a few object categories (such as cars) make both across-class and within-class recognition necessary. Each object is depicted rotated in depth from its original position in increments of 5 degrees. We chose the 0 degree images of each object as training images, and used the following 9 images as test images. The only preprocessing performed on these images was reducing them from full color to grayscale.

3.2 Procedure. To determine the relative performance of center-surround, local-oriented, and nonlocal features in an unbiased way, we model all of our features as generalized difference-of-gaussian operators. A generic bilobed operator in two-dimensional space can be modeled as follows:

$$\frac{1}{\sqrt{2\pi}\,|\Sigma_1|^{1/2}}e^{\frac{-(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)}{2}} - \frac{1}{\sqrt{2\pi}|\Sigma_2|^{1/2}}e^{\frac{-(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2)}{2}}. \tag{3.1}$$

For all of our remaining experiments, we consider only operators with diagonal covariance matrices Σ_1 and Σ_2. Further, the diagonal elements of each matrix Σ shall be equal, yielding isotropic gaussian lobes. For this simplified case, equation 3.1 can be expressed as

$$\frac{1}{\sqrt{2\pi}\sigma_1}e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} - \frac{1}{\sqrt{2\pi}\sigma_2}e^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}}. \tag{3.2}$$

We introduce also a parameter δ to represent the separation between two lobes. This is simply the Euclidean norm of the difference between the two means:

$$\delta = \|\mu_2 - \mu_1\|. \tag{3.3}$$

In order to build a center-surround operator, δ must be set to zero, and the spatial constants of the center and surround should be in a ratio of 1 to 1.6 to match the dimensions of receptive fields found in the human visual system (Marr, 1982). To create a local-oriented operator, we shall set $\sigma 1 = \sigma 2$, and set the distance δ to be equal to three times the value of the spatial constant. Finally, nonlocal operators can be created by allowing the distance δ to exceed the value 3σ (once again assuming equal spatial constants for the two lobes). Examples of all of these operators are displayed in Figure 6.

Given this simple parameterization of our three feature types, we choose in this experiment to sample equal numbers of each kind of operator from the full set of possible features. In this way, we may represent each of our training images in terms of some small number of features drawn from a specific operator family and evaluate subsequent classification performance.

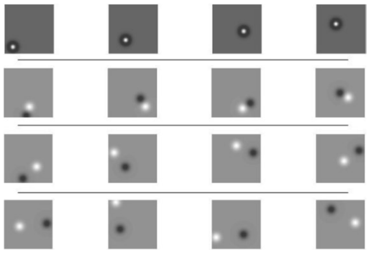

Figure 6: Representative operators drawn from the four operator families considered in experiment 2. Top to bottom, we display examples of center-surround features, local oriented features, and two kinds of nonlocal features ($\delta = 6\sigma, s = 9\sigma$).

Four operator families were considered: center-surround features ($\delta = 0$), local-oriented features ($\delta = 3\sigma$), and two kinds of nonlocal features ($\delta = 6\sigma$ and 9σ). For each operator family, we constructed 40 banks of 50 randomly positioned and oriented operators each. Twenty of these feature banks contained operators with a spatial constant of 2 pixels, and the other 20 feature banks contained operators with a 4 pixel spatial constant. Each bank of operators was applied to the training images to generate a feature vector consisting of 50 values. The same operators were then applied to all test images, and the resulting feature vectors were classified using a nearest-neighbor metric (L2 norm). This procedure was carried out on both the ORL and the COIL databases.

3.3 Results. The number of images correctly identified for a given filter bank was calculated for each recognition trial, allowing us to compute an average level of classification performance from the 20 runs within each operator family and spatial scale (see Figure 7). We find in this task that once again, center-surround and nonlocal features offer the best recognition performance. This result holds at both spatial scales used in this task, as well as for both face recognition and multiclass object recognition. We also note the small variability in recognition performance around each operator's mean value. Despite the random sampling of features used to constitute our operator banks, the resulting recognition performance remained very consistent.

In both cases, we note that center-surround performance slightly exceeds that obtained using nonlocal operators. It is interesting to note, however, that a larger separation between the lobes of a nonlocal feature results in better recognition performance. This cannot continue indefinitely, of course, as longer and longer separations will lead to more limitations on where operators can be placed within the image. Increased accuracy with increased

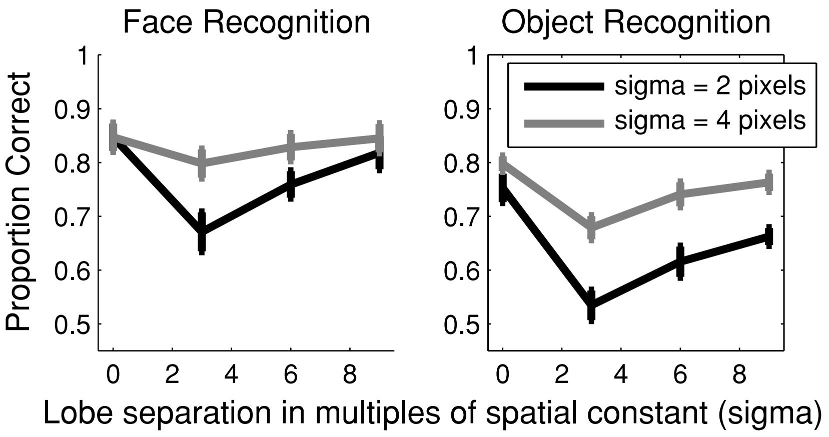

Figure 7: Recognition performance for both faces (left) and objects (right) as a function of both the distance between operator lobes and the spatial constant of the lobes.

nonlocality does suggest that larger distances between lobes are more useful, however, and that it is not enough simply to deviate from locality.

We note that the distinct dip in performance for local-oriented features is both consistent and puzzling. Why should it be the case that unoriented local features are good at recognition while oriented local features are poor? Center-surround operators analyze almost the same pixels as a local-oriented operator placed at the same location, so why should they be so different in terms of their recognition performance? Moreover, how is it that radically different operators like the dissociated dipole and the center-surround operator should perform so similarly? In our third and final experiment, we attempt to address these questions by breaking down the recognition problem into distinct parts so we can learn how these operator families function in classification tasks.

Specifically, we point out that good recognition performance is made possible when an operator possesses two distinct properties. First, an operator must provide a stable response to images of objects with the same identity. Second, the operator must respond differently to images of objects with different identities. Neither condition is sufficient for recognition to proceed, but both are necessary. We hypothesize that though both center-surround operators and nonlocal operators provide useful information for recognition, they do so in different ways. In our last experiment, we assess both the stability and variability of each operator type to determine how good recognition results are achieved with different receptive field structures.

4 Experiment 3: Feature Stability and Variability

In experiment 2, we determined that both center-surround and nonlocal operators outperform local oriented features at recognition of faces and objects. In many ways, this is quite surprising. Center-surround features appear to share little with nonlocal operators as we have defined them, yet their recognition performance is quite similar.

In this task, we break down the recognition process into components of stability and variability. To perform well at recognition, a particular operator must first be able to respond in much the same way to many different images of the same face. This is how we define stability, and one can think of it in terms of various identity-preserving transformations. Whether a face is smiling or not, lit from the side or not, a useful operator for recognition must not vary its response too widely. If this proves true, we may say that that feature is stable with respect to the transformation being considered.

We use this notion to formulate an operational definition of stability in terms of a set of image measurements and a particular face transformation. Let us first assume that we possess a set of image measurements in a filter bank, just as we did in experiment 2. This filter bank is applied to some initial image, which shall always depict a person in frontal view with a neutral expression. The value of each operator in our collection can be determined and stored in a one-dimensional vector, x. This same set of operators is then applied to a second image, depicting the same person as the original image but with some change of expression or pose. The values resulting from applying all operators to this new image are then stored in a second vector, y. The two vectors x and y may then be compared to see how drastic the changes in operator response were across the transformation from the first image to the second. If by some luck our operators are perfectly invariant to the current transformation, plotting x versus y would produce a scatter plot in which all points would lie on the line $y = x$. Poor invariance would be reflected in a plot in which points are distributed randomly. For two vectors x and y (each of length n), we may use the value of the correlation coefficient (see equation 4.1) between them as our quantitative measure of feature stability:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}. \tag{4.1}$$

The second component of recognition is variability. It is not enough to be stable to transformations; one must also be diagnostic of identity. Imagine, for example, that one finds an image measurement that is perfectly stable across lighting, expression, and pose transformations. It may seem that this measurement is ideal for recognition, but let us also imagine that it turns out to be of the same value for every face considered. This provides no means of distinguishing one face from another, despite the measurement's remarkable invariance to transformations of a single face. What is needed is an ability to be stable within images of a single face, but vary broadly across images of many different faces. This last attribute we shall call variability, and we may quantify it for a particular measurement as the variance of its response across a population of faces.

In this third experiment, we use these operational definitions of stability and variability to determine what properties center-surround and nonlocal operators possess that make them useful for recognition. We shall return once again to the domain of faces, as they provide a rich set of transformations to consider, both rigid and nonrigid alterations of the face in varying degree.

4.1 Stimuli. We use 16 faces (8 men, 8 women) from the Stirling face database for this experiment. The faces are grayscale images of individuals in a neutral, frontal pose accompanied by pictures of the same models smiling and speaking while facing forward, and also in a three-quarter pose with neutral expression. We call these transformations the SMILE, SPEECH, and VIEW transforms, respectively. The original images were 284 × 365 pixels, and the only preprocessing step applied was to crop out a 256 × 256 pixel region centered in the original image rectangle.

4.2 Procedure. All operators in these sets were built as difference-of-gaussian features, exactly as described in experiment 2. Also as before, center-surround, local oriented, and two kinds of nonlocal features were evaluated. Because we would like to understand how both the separation of lobes and their individual spatial extent affect performance, two scales were employed for each kind of feature. Space constants of 4 pixels (fine scale) and 8 pixels (coarse scale) were used. In the case of center-surround features, the value of the space constant always refers to the size of the surround. For each pair of images to be analyzed, we construct 120 collections of 50 operators each. These feature banks were split into 10 center-surround, 10 local, and 20 nonlocal banks (10 banks each for separations of six and nine times the spatial constant of the lobes) at both scales mentioned above.

Once a set of operators was constructed, we applied it to each neutral, frontal image in our data set to assemble the feature value for the starting image. The same operators were then applied to each of the three transformed images so that a value for Pearson's *R* could be calculated for that set of operators relative to each transformation. The average value of Pearson's *R* could then be taken across all 16 faces in our set. This process was repeated for all families and scales of operator banks to assess stability.

To assess variability, operator banks were once again applied to the neutral, frontal images once again. This time, the variance in each operator's output was calculated across the population of 16 faces. The results were

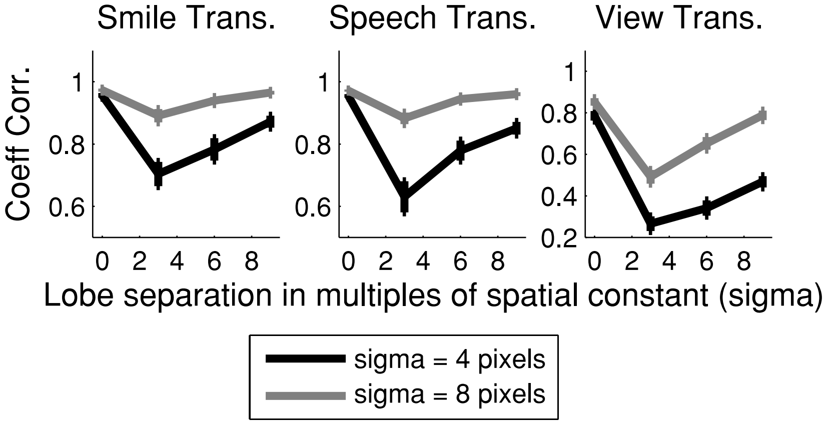

Figure 8: The stability of each feature type (x-axis) as a function of both the spatial scale of the gaussian lobes and various facial transformations.

combined and expressed in terms of the mean variance of response and its standard deviation.

4.3 Results

4.3.1 Difference-of-Gaussian Features. Plots depicting the average values of the correlation coefficients (averaged again over all individuals) are presented in Figure 8. We present the measured stability of each kind of operator across three ecologically relevant transformations: SMILE (second image of individuals smiling), SPEECH (second image of individuals speaking), and VIEW (second image of individuals in three-quarters pose).

These plots highlight several interesting characteristics of our operators. First, center-surround filters at both scales appear to perform quite well compared to the other features once again. As soon as we move the two gaussians apart to form oriented local operators, however, a sharp dip in stability occurs. This indicates that the two-lobed oriented edge detectors used here provide for comparatively poor stability across all three of the transformations we have examined here. That said, as the distance between the lobes of our operators increases further, stability of response also increases. Nonlocality seems to increase stability across all three transformations, nearly reaching the level of center-surround stability at a coarse scale.

Stability, however, is not the only attribute required to perform recognition tasks well. As discussed earlier, a feature that is stable across face transformations is useful only if it is not also stable across images of different individuals. That is, a universal feature is not of any use for recognition

Table 1: Mean ± S.E. of Operator Variance Across Individuals.

	$\sigma = 4$	$\sigma = 8$	$\sigma = 16$
Center-surround	122.5 ± 3.7	206.6 ± 6.2	311.3 ± 8.5
Local (s = 3)	242.0 ± 9.6	527.0 ± 15.0	986.9 ± 26.7
Nonlocal (s = 6)	378.8 ± 11.4	718.5 ± 17.7	1204.1 ± 29.9
Nonlocal (s = 9)	430.2 ± 11.0	795.4 ± 19.7	1271.7 ± 32.6

because it has no discriminative power. We present next the amount of variability in response for each family of operators (see Table 1).

Center-surround operators appear to be the least variable across images of different individuals, while nonlocal operators appear to vary most. All feature types except for the center-surround filters increase in variability as their scale increases, which seems somewhat surprising, as one might expect more dramatic differences in individual appearance to be expressed at a finer scale. Nonetheless, we can see from the combination of these results and the stability results that center-surround and nonlocal operators achieve good recognition performance through different means. Center-surround operators are not so variable from person to person, but make up for it with an extremely stable response to individual faces despite significant transformations. In contrast, nonlocal operators lack the full stability of center-surround operators, but appear to make up for it by being much more variable in response across the population of faces. The local-oriented features rank poorly in terms of both their stability and variability characteristics, thus limiting their usefulness for recognition tasks.

4.4 Discussion. The results of our stability analysis of differential operators reveal two main findings. First, the same features that were discovered to perform the best discrimination between intra- and interpersonal difference vectors in experiment 1 (large center-surround filters and nonlocal operators) and to perform best in a simple recognition system for both faces and objects (experiment 2) also display the greatest combination of stability and variability when confronted with ecologically relevant face transforms. However, the limited stability of local oriented operators suggests that they may not provide the most useful features for handling these image transforms.

5 Conclusion

We have noted the emergence of large center-surround and nonlocal operators as tools for performing object recognition using simple features and found that both of these operators provide good stability of response across a range of different transforms. These structures differ from receptive field forms known to support sparse encoding of natural scenes, yet seem to provide a better means of discriminating between individual objects and providing stable responses to image transforms. This suggests that the constraints that govern information-theoretic approaches to image representation may not necessarily be useful for developing representations that can support the recognition of objects in images.

In the specific context of faces, do large center-surround fields or nonlocal comparators, on their own, present a viable alternative to performing efficient face recognition? At present, the answer to this question is no. Complex (and truly global) features such as eigenface (Turk & Pentland, 1991) bases provide for higher levels of recognition performance than we expect to achieve using these far simpler features. We note, however, that the discovery of a useful vocabulary of low-level features may aid global recognition techniques like eigenface-based systems. One could easily compute PCA bases on nonlocal and center-surround measurements rather than pixels. The added stability of these operators may help significantly increase recognition performance.

The larger question at stake, however, does not only concern face recognition, despite its' being our domain of choice for this study. Of greater interest than building a face recognition engine is learning how one might obtain stability to relevant image transforms given some set of simple measures. Little is known about how one moves from highly selective, small receptive fields in V1 to the large receptive fields in inferotemporal cortex that demonstrate impressive invariance to stimulus manipulations within a particular class. We have introduced here a particular measurement, the dissociated dipole, which represents one example of a very broad space of alternative computations by which limited amounts of invariance might be achieved. Our proposal of nonlocal operators draws support from several studies of human perception. Indeed, past psychophysical studies of the long-range processing of pairs of lines suggest the existence of similarly structured "coincidence detectors," which enact non-local comparisons of simple stimuli (Morgan & Regan, 1987; Kohly & Regan, 2000). Further work exploring nonlocal processing of orientation and contrast has more recently given rise to the idea of a "cerebral bus" shuttling information between distant points (Danilova & Mollon, 2003). These detectors could contribute to shape representation, as demonstrated by Burbeck's idea of encoding shapes via medial "cores" built by integrating information across disparate "boundariness" detectors (Burbeck & Pizer, 1995).

Our overarching goal in this work is to redirect the study of nonclassical receptive field structures toward examining the possibility that object recognition may be governed by computations outside the realm of traditional multiscale pyramids, and subject to different constraints from those that guide formulations of image representation based on information theory. The road from V1 to IT (and, computationally speaking, from Gabors and gaussian derivatives to eigenfaces) may contain many surprising image processing tools.

Even within the realm of dissociated dipoles, there are many parameters to explore. For example, the two lobes need not be isotropic or be of equal size and orientation. The lobes could easily take the form of gaussian derivatives rather than gaussians. Given that there are many more parameters that could be introduced to the simple DOG framework, it is possible that even better invariance could be achieved by introducing more degrees of structural freedom. The point is that expanding our consideration to nonlocal operators opens up a large space of possible filters, and systematic exploration of this space, while difficult, may be very rewarding.

Acknowledgments

This research was funded in part by the DARPA HumanID Program and the National Science Foundation ITR Program. B.J.B. is supported by an NDSEG fellowship. P.S. is supported by an Alfred P. Sloan Fellowship in neuroscience and a Merck Foundation Fellowship. We also thank Ted Adelson, Gadi Geiger, Mriganka Sur, Chris Moore, Erin Conwell, and David Cox for many helpful discussions.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network, 3*, 213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev., 61*, 183–193.

Balas, B. J., & Sinha, P. (2003). *Dissociated dipoles: Image representation via nonlocal operators*. Cambridge, MA: MIT Press.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication.* Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Borenstein, E., & Ullman, S. (2002). Class-specific, top-down segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 109–124). Berlin: Spring-Verlag.

Borenstein, E., & Ullman, S. (2004). Learning to segment. In *Proceedings of the European Conference on Computer Vision* (pp. 315–328). Berlin: Springer-Verlag.

Burbeck, C. A., & Pizer, S. M. (1995). Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research, 35*(13), 1917–1930.

Chapin, J. K. (1986). Laminar differences in sizes, shapes, and response profiles of cutaneous receptive fields in the rat SI cortex. *Exp. Brain Research, 62*(3), 549–559.

Danilova, M. V., & Mollon, J. D. (2003). Comparison at a distance. *Perception, 32*(4), 395–414.

Edelman, S. (1993). Representing 3-D objects by sets of activities of receptive fields. *Biological Cybernetics, 70*, 37–45.

Fei-Fei, L., Fergus, R., & Perona, P. (2003). *A Bayesian approach to unsupervised one-shot learning of object categories*. Paper presented at the International Conference on Computer Vision, Nice, France.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). *Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories*. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6*, 559–601.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(9), 891–906.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology, 148*, 574–591.

Julesz, B. (1975). Experiments in the visual perception of texture. *Scientific American, 232*(4), 34–43.

Kersten, D. (1987). Predictability and redundancy of natural images. *J. Opt. Soc. Am. A, 4*(12), 2395–2400.

Kohly, R. P., & Regan, D. (2000). Coincidence detectors: Visual processing of a pair of lines and implications for shape discrimination. *Vision Research, 40*(17), 2291–2306.

Kouh, M., & Riesenhuber, M. (2003). *Investigating shape representation in area V4 with HMAX: Orientation and grating selectivities*. (Rep. AIM=2003=021, CBCL=231). Cambridge, MA: MIT.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch, 36*, 910–912.

Marr, D. (1982) *Vision*. New York: Freeman.

Moghaddam, B., Jebara, T., & Pentland, A. (2000). Bayesian face recognition. *Pattern Recognition, 333*(11), 1771–1782.

Morgan, M. J., & Regan, D. (1987). Opponent model for line interval discrimination: Interval and Vernier performance compared. *Vision Research, 27*(1), 107–118.

Nayar, S. K., Nene, S. A., & Murase, H. (1996). *Real-time 100 object recognition system*. Paper presented at the ARPA Image Understanding Workshop, Palm Springs, FL.

Nene, S. A., Nayar, S. K., & Murase, H. (1996). *Columbia Object Image Library* (*COIL-100*). New York: New York University.

Novikoff, A. (1962). Integral geometry as a tool in pattern perception. In H. Foerster & G. Zopf (Eds.), *Principles of self-organization*. New York: Pergamon.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Samaria, F., & Harter, A. (1994). *Parametrisation of a stochastic model for human face identification*. Paper presented at the Second IEEE Workshop on Applications of Computer Vision, Sarasota, FL.

Schneiderman, H., & Kanade, T. (1998). *Probabilistic modeling of local appearance and spatial relationships for object recognition*. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Sinha, P. (2002). Qualitative representations for recognition. *Lecture Notes in Computer Science, 2525*, 249–262.

Turk, M. A., & Pentland, A. P. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI.

Wiskott, L., Fellous, J.-M., Kruger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 775–779.

Young, E. D. (1984). Response characteristics of neurons of the cochlear nucleus. In C. I. Berlin (Ed.), *Hearing science recent advances*. San Diego, CA: College Hill Press.

Received July 28, 2004; accepted August 9, 2005.