Pergamon

PII: S0042-6989(97)00187-9

Vision Res., Vol. 37, No. 23, pp. 3343–3365, 1997

Printed in Great Britain
0042-6989/97 $17.00 + 0.00

Structure Detection: A Statistically Certified Unsupervised Learning Procedure

CHARLES CHUBB,*† ZHONG-LIN LU,‡ GEORGE SPERLING†

Received 16 July 1996; in revised form 25 March 1997; in final form 11 June 1997

We present a class of structure detection procedures (SDPs) that can extract the characteristic structures in an arbitrary population of images. An SDP adaptively augments the power of a novel, statistical, structure test to reject the null hypothesis that a randomly chosen image is devoid of structure. The core of the structure test consists of an orthonormal basis B of receptive fields that is refined into an increasingly sensitive detector of characteristic image structures. Adaptive refinement is accomplished as follows: for each image x in a random training sequence, B is updated by a planar rotation that decreases the p-value of a statistical structure test for x. This image-by-image refinement procedure is very efficient, obeying time and space constraints similar to those that limit processes of perceptual organization in real organisms. SDPs' capabilities are demonstrated in three test populations: natural images, faulty random number generators, and artificial images composed of mixtures of basis functions. (1) An SDP succeeds in rejecting the null hypothesis that the UNIX random number generator rand() is truly random. (2) When images are composed by adding arbitrary pairs of orthogonal component images, an SDP extracts the components. (3) For a large set of natural image patches, an SDP yields a basis B_1 that detects structure with p-value < 0.005 in 88% of a new set of patches. B_1's elements resemble the receptive fields of V1 simple cells. (4) Of special interest are biconvergent SDPs that derive in parallel a basis B, as well as a pointwise transformation f, specifically sensitized to evaluate the response values that result from applying B to images in the target population. A biconvergent SDP applied to natural image patches yields a basis B_2 similar to B_1, as well as a pointwise transformation f with vastly heightened sensitivity to extreme response values. We conjecture that sensory neurons have evolved cooperatively to maximize their collective power to reject the null hypothesis that their input is devoid of structure, thereby evolving receptive fields that efficiently represent characteristic input structures. © 1997 Elsevier Science Ltd

Natural images Neural networks Structure Unsupervised learning Image coding

INTRODUCTION

It is now well established that the visual system has a hierarchical organization. We propose that this hierarchical organization reflects a general strategy, achieved through evolution, for detecting environmental regularities at different levels of abstraction. Specifically, we suggest that in each processing stage, the goal is to detect the characteristic structures in the input—the output from the previous stage. In this paper, we describe a procedure that can be used to detect input structures. We speculate that the emergent structures at various levels of human visual processing may reflect the recursive application of such a procedure.

Structureless worlds

Imagine a world in which visual input is totally devoid of structure. One such world is completely without light. Many species that live in the ocean depths adapt to such a world, and generally they evolve without sight. An equally structureless, hypothetical world is one that presents only spatially uniform, temporally unvarying light of an unchanging spectral composition.

Visual input also is devoid of structure in a world that presents to the retina a constant storm of homogeneous white noise. In such a world, the intensity stimulating any given point in the retina at a given time is perfectly unpredictable; aside from measuring it directly, there is no observation, no experiment, that an observer could make that would provide any purchase whatsoever in guessing this intensity. This means that there is no possibility of accounting for the images presented to the

*To whom all correspondence should be addressed [*Tel*: (714) 824-3517; *Fax*: (714) 824-2517; *Email*: cfchubb@uci.edu].

†Department of Cognitive Sciences, Institute for Mathematical Behavioral Sciences, University of California at Irvine, Irvine, CA 92697, U.S.A.

‡Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061, U.S.A.

retina in terms of any construct more compact than the entire image stream itself. Vision would be just as useless in this white noise world as it would be in a world of completely uniform stimulation.

These remarks presuppose an intuitive understanding of the term 'structure' as it applies to images and image populations. The purpose of the rest of this section is to give the term "structure" a more precise, formal foundation.

Images, random images and image populations

We consider image populations, each of whose images comprises some fixed number N of pixels. An image v can be considered as a vector in $\mathbb{R}^N$. The i^{th} coordinate value of v is thought of as the i^{th} pixel value of v and is denoted $v[i]$. The term "random image" is often used to refer to an image whose pixel intensities are randomly assigned; here, however, a *random image* x is simply a random variable in $\mathbb{R}^N$ (with $\mathbb{R}^N$ now construed as the set of all images). That is, x is a random selection from a population of images characterized by a density f on the set $\mathbb{R}^N$ of images.

Structureless random images and image populations

Some random images have the property that any image that results from scrambling their pixels has the same probability as the original image. We capture this notion formally as follows: let f be a probability density on $\mathbb{R}^N$. A given image v is said to be *scramblable under f* if $f(q) = f(v)$ for any image q whose pixel-intensity histogram is identical to that of v.

A random image x with probability density f is said to be *structureless* if every image is scramblable under f. This means that the probability that x is equal to a given image v depends only on the pixel-intensity histogram of v, and is completely independent of the arrangement of the pixels.

Structureless does not mean Useless

Note the curious implication that any uniform image v (i.e., any v whose pixels all take the same value) is scramblable under *any* probability density f. Thus, any random image x is structureless if its density assigns non-zero values only to uniform images.

A very simple structureless random image, x, is either all white or else all black with equal probability. A world populated by such images would be either all white or all black at any given instant. It might well be important for an animal or a microbe in such a world to be able to discriminate the "white world" from the "black world" if, for instance, the white world affords different behavioral possibilities than the black world.

This example illustrates that rudimentary visual processes might be useful in worlds with structureless image populations. However, such worlds are devoid of all explicitly *spatial* structure; vision is effectively reduced to a purely temporal sense, akin to smell.

The unique potential of vision, however, derives from its sensitivity to forms and patterns, to relations between intensities occurring at different locations in space. This paper focuses exclusively on such spatial relations.

Latent structures and projection pursuit

How does vision, or indeed any sensory system, without guidance, become sensitized to the characteristic structures in its world?

In this paper, we investigate the possibility that vision achieves its environment-specific sensitivity by adaptively increasing its power to statistically reject the null hypothesis H_0 that its input is structureless. A critical principle is that precisely the same procedure can be used at the next level of processing, and indeed at successively higher levels to discover the higher order, latent structures in the visual input.

We describe adaptive processes called structure detection procedures, SDPs, that can be used to test H_0. We show that, when applied to the population of natural images, such a process naturally generates a set of receptive fields that resemble simple cell receptive fields.

This project falls into a large body of recent research devoted to understanding the relationship between the statistics of natural images and the structure of simple cell receptive fields (e.g., Barrow, 1987; Barlow, 1989; Law & Cooper, 1994; Fyfe & Baddeley, 1995; Schmidhuber, Eldracher, & Foltin, 1996; Linsker, 1988; Oja, 1989; Sanger, 1989; Olshausen & Field, 1996; Harpur & Prager, 1996; Foldiak, 1990; Intrator & Cooper, 1992; Intrator, 1992; Hancock, Baddeley, & Smith, 1992; Bell & Sejnowski, 1995; Liu & Shouval, 1994; Ruderman & Bialek, 1994; Ruderman & Bialek, 1992; Shouval & Liu, 1996). We do not claim biological plausibility for the computation to be described. Our aim here is to understand the goal of visual recoding, not necessarily the biological process by which that goal is achieved.

Projection pursuit. Projection pursuit (e.g., Huber, 1985) is a general method used in searching for structure in complicated data sets. The relation of structure detection procedures to projection pursuit is an interesting issue that will be considered in detail after SDPs have been described.

Overview of a structure detection procedure (Fig. 1)

Matrix of weights, B. The core of the SDP is a matrix B that can be considered as the current set of receptive fields used to process the image. The B receptive fields are orthonormal, and initially chosen arbitrarily to begin the SDP.

Gaussian replacement. An image x is then chosen from the population. Images have N pixels; the matrix B is $N \times N$. The intensities of x's pixels are ordered from lowest to highest, and then replaced by the correspondingly ordered values in an independent sample of standard Normal random variables. The resulting image, $G(x)$ is called the *gaussian replacement* of x. The histogram of $G(x)$'s pixel values is thus precisely the histogram of a sample of standard normal random variables. The gaussian replacement transformation is

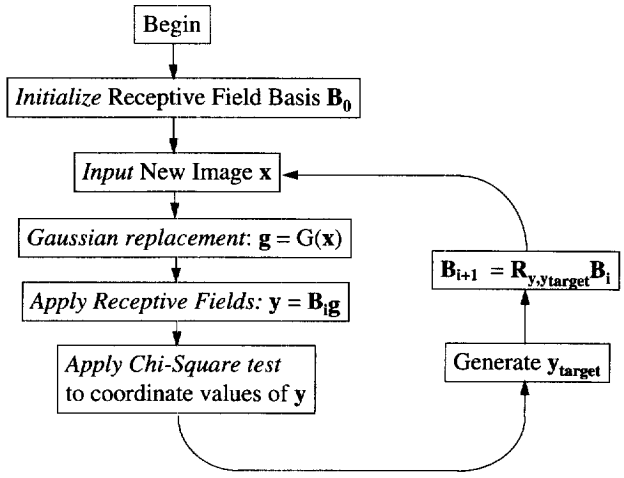


FIGURE 1. The flow of a Structure detection procedure. Following random initialization of the $N \times N$ basis B_0 of receptive fields, the SDP enters a loop, each iteration of which constitutes a single training trial. In the $i + 1^{st}$ training trial, a random image x is drawn from the target image population. The input image x is then transformed into a realization g of $G(x)$, the gaussian replacement of x. Next, the current basis B_i is applied to g yielding a vector y of receptive field responses. Under H_0, y should consist of jointly independent, standard normal random variables. A chi-square test is applied to the coordinate values of y in an attempt to reject H_0. Then a vector y_{target} is selected in the neighborhood of y such that (i) the histogram of y_{target} differs more from normality than that of y; and (ii) $|y_{target}| = |y|$ (so that y_{target} is reachable from y by a rotation). Then, for $R_{y,y_{target}}$ the planar rotation in the plane spanned by $\{y, y_{target}\}$ that maps y onto y_{target}, B_{i+1} is set equal to $R_{y,y_{target}} B_i$. This assignment has the desired effect that $B_{i+1}g = y_{target}$; thus, applying B_{i+1} to g yields a vector y_{target} of receptive field responses that deviate from normality more than the receptive field responses $y = B_i g$.

useful for statistical certification of the SDP, but is not essential for other aspects of the SDP, such as arriving at efficient representations of the input.

Linear transformations. The matrix B of receptive fields is applied to $G(x)$ yielding a transformed image $B(G(x))$. The histogram of $B(G(x))$ is evaluated for Normality, and a planar rotation is applied to B to produce a new basis B', such that $B'(G(x))$ deviates slightly more from Normality than $B(G(x))$. Then the process is repeated with a new image.

The use of a planar rotation to update B permits a significant simplification and speedup of the adaptive search process.

Statistical certification. Structure detection is *statistically certified* in the following sense: the set B of receptive fields derived as the end product from applying structure detection to a sample drawn from an image population P is a P-specific, statistical tool that can be used on any new image x of P to test the null hypothesis H_0 that x is structureless. If and only if the image population P is structured, is it possible to derive a set B empowered to reject H_0 with probability greater than chance.

However, statistical certification is purchased with some costs. (1) The logic underlying structure detection requires that the emergent basis B of receptive fields be orthonormal. This condition is not required by most unsupervised learning procedures; moreover, it has been explicitly argued (Olshausen & Field, 1996a) that the statistics of natural images make it unlikely (and counterproductive) for simple cell receptive fields to strive for orthonormality. As we shall explain, though, constraining B to be orthonormal confers important inferential power. (2) The current procedure involves a preprocessing stage that discards a great deal of information from input images. In particular, the intensities of individual pixels are ordered from lowest to highest, and then replaced by the correspondingly ordered values in a sample of standard Normal random variables. This operation preserves only the ordinal information in the input image. (3) The initial image transformation involves randomness which makes it noninvertible.

The section entitled "*Structure detection*" below gives a precise description of the structure detection procedure. The description of the procedure is given in "*The tuning procedure*". In "*Preliminary tests of structure detection*", we test the procedure on two image populations: images consisting of bit-strings from the Unix random number generator rand(), and one other artificial population of images. Then in "*Discovering structure in natural images with an SDP*" we describe an application of structure detection to the population of natural images. The resulting receptive fields are compared with simple cells. "*A search procedure without gaussian replacement*" investigates the importance of gaussian replacement in the search procedure of an SDP; specifically, an SDP without gaussian replacement is applied to natural images. "*Biconvergent SDPs*" are introduced in the following section. These SDPs simultaneously discover a basis of receptive fields as well as the way in which their responses deviate from normality. No artificial update rule is imposed to guide the search procedure. Finally, "*A simulation using a biconvergent*

SDP" reports the result of a biconvergent SDP applied to natural images.

STRUCTURE DETECTION

In this section we describe *structure detection*, an adaptive procedure for discovering the characteristic structures inherent in a given image population. We begin with several preliminary definitions.

Preliminary definitions

The results we present here presuppose that random variables are real-valued and continuous.

A function T: $\mathbb{R}^N \to \mathbb{R}^N$ is called an image transformation. For any image v, $T(v)$ denotes the image that results from applying T to v, and $T(v)[i]$ gives the i^{th} pixel value of $T(v)$.

Standard normal IID random images. A random image x is called IID if x's pixel values are jointly independent, identically distributed random variables. In this case, the (cumulative) distribution function characterizing one of x's pixel values is called x's pixel distribution. An IID random image x is called standard Normal if its pixel values are standard normal random variables.

Pointwise transformations and histogram distortion templates. For any function f:$\mathbb{R} \to \mathbb{R}$, and any image v, we define the image $f \circ v$ by setting

$$(f \circ v)[j] = f(v[j]) \tag{1}$$

for all pixels $j = 1, 2, \ldots, N$. $f\circ$ is thus a transformation whose output value at any pixel results from applying the function f to the input value at that pixel. Accordingly, $f \circ$ is called a pointwise transformation.

For reasons that will become clear later, we call a pointwise transformation $f\circ$ a histogram distortion template if

$$\int_{-1}^{1} f(r)dr = 0 \tag{2a}$$

$$\int_{-1}^{1} f^2(r)dr = 1. \tag{2b}$$

Isometries. Let B be an $N \times N$ matrix whose row vectors make up an orthonormal basis of $\mathbb{R}^N$. In this case, B's inverse is B^T (the transpose of B). Thus, for any $v \epsilon \mathbb{R}^N$,

$$|Bv|^2 = (Bv)^{\mathrm{T}}(Bv) = v^{\mathrm{T}}B^{\mathrm{T}}Bv = v^{\mathrm{T}}v = |v|^2, \tag{3}$$

showing that vector length is preserved by the linear operator B; for this reason B is called an isometry.

Two-dimensional (planar) rotations. For any distinct, orthonormal vectors $x, y \in \mathbb{R}^N$, any $\beta \in [0, 1]$, the matrix

$$Q_{x,y,\beta} = I + \left[x(\beta - 1) + y\sqrt{1 - \beta^2}\right]x^{\mathrm{T}} + \left[y(\beta - 1) - x\sqrt{1 - \beta^2}\right]y^{\mathrm{T}} \tag{4}$$

is an isometry. In particular, for any $v \epsilon \mathbb{R}^N$, $Q_{x,y,\beta}v$ is the vector that results from rotating the projection of v in the plane spanned by $\{x, y\}$ through an angle $\theta = arccos(\beta)$ in the direction from x toward y.

This leads to the following definition. For arbitrary normal vectors $x, y \in \mathbb{R}^N$ (x and y not necessarily orthogonal), set

$$R_{x,y} = Q_{x,\bar{y},x\cdot y}, \tag{5}$$

where $\bar{y}$ is the normalized component of y orthogonal to x:

$$\bar{y} = \frac{y - (x \cdot y)x}{|y - (x \cdot y)x|}. \tag{6}$$

It is easy to check that:

$$R_{x,y}x = y. \tag{7}$$

Thus, $R_{x,y}$ is the rotation within the plane spanned by $\{x, y\}$ that maps x onto y.

The structure test

Structure detection is an adaptive method for tuning an isometry B to detect the characteristic structures in a population P of images. At the core of this method is the statistical test that we call the structure test. The structure test depends on the following well-known fact.

Observation 1. An isometry applied to an IID standard Normal random image yields a IID standard Normal random image. That is, for any isometry B on $\mathbb{R}^N$, and any standard Normal IID image y, By is also a standard Normal IID image. The proof (omitted for brevity) depends on the fact that the joint density of a standard Normal IID image is spherically symmetric, and hence will be preserved under rotations.

We make use of this fact as follows. Let x be a random image from P, and let B be an orthonormal basis spanning $\mathbb{R}^N$. Assume the null hypothesis H_0 that x is devoid of structure. We perform the following steps:

1. Gaussian replacement
2. B-application
3. Normality testing

Gaussian replacement transforms the random image x into a random image $G(x)$ with the following property: if x is structureless, then $G(x)$ will be standard Normal IID. Hence, by the observation above, the result $B(G(x))$ of applying isometry B to $G(x)$ (step 2) must also be standard Normal IID. We test this condition in step 3; in particular, we use a standard chi-square test of the hypothesis that the histogram of $B(G(x))$ was generated by a sample of N jointly independent, standard Normal random variables. We now describe these three steps in detail.

Gaussian replacement. The goal of this step is to derive from the given image x a random image $G(x)$ that will be standard Normal IID if (and only if) x is structureless. To do this we

1. Obtain a fresh sample S of N jointly independent standard Normal random variables. Then,
2. Produce $G(x)$ by replacing the i^{th} greatest pixel-value of x (for $i = 1, 2, \ldots, N$) by the i^{th} greatest value in the random sample S. In the case in which the i^{th}, $i + 1^{st}, \ldots, i + k^{th}$ greatest pixel-values of x are all

equal, the corresponding pixels of $G(x)$ are assigned the i^{th}, $i+1^{st}$,..., $i+k^{th}$ greatest values of S in random order.

The Gaussian replacement operator G has some unusual properties. First, note that G is not an ordinary image transformation; it is a random image transformation. That is, given a non-random input image $v \epsilon \mathbb{R}^N$, $G(v)$ is a random image. However, G is quite well-behaved for images v with many distinct pixel intensities. In particular, the larger we make the number of pixels in an image, the less randomness G introduces when applied to richly variable images.

Note also that $G(v)$ depends only on the ordinal relations between the pixel intensities in v. Thus, for any images $v, w \epsilon \mathbb{R}^N$, if the ordinal relations among the pixel intensities of v are identical to the ordinal relations among the pixel intensities of w, then $G(v)$ is identically distributed to $G(w)$.

This is potentially a useful property for a visual preprocessing transformation to have. Vision is primarily concerned with extracting information about things in the world. However, the light emanating from objects depends both on the surface properties of the objects and also on the spectral properties of the illuminating light. Accordingly, as many have noted, a plausible goal of early visual processing is to transform the retinal image so as to discard information about the illuminant, while preserving only information about illuminated objects. Of course, the absolute light levels of points in the visual field are liable to depend in uncontrolled ways on the nature of the illuminant. Conversely, *ordinal* relations between intensities in the scene tend to be invariant with respect to variations in illumination over time.

For purposes of the structure test, the important point to note about G is that if the image x being tested is structureless (as assumed under H_0), then, $G(x)$ will be standard Normal IID. On the other hand, if x is not structureless, then the ordinal, interpixel relations within x will be largely preserved in $G(x)$.

B-application. Next we apply B to $G(x)$. If H_0 holds, then the observation above implies that $B(G(x))$ must be a standard Normal IID random image. On the other hand, if H_0 is false then $G(x)$'s pixel values will be systematically ordered across space. In this case, applying an appropriate basis B to $G(x)$ yields a set of response values whose histogram deviates significantly from standard normality.

Normality testing. If we can reject the hypothesis that the response values of $B(G(x))$ are jointly independent standard Normal random variables, we also reject the hypothesis that $G(x)$ is a standard Normal IID function, which in turn rejects the original null hypothesis H_0 that x is structureless.

In the first SDPs we shall describe, the test we use (others would have served as well) is a standard chi-square test in which we partition the real number line into bins $C_1, C_2,..., C_m$ subsuming equal area (*bin area* = $1/m$) under the standard Normal density curve. Then, under H_0, the expected number of $B(G(x))$ values within each bin is N/m. Suppose the actual number of observations falling within bin i is $O[i]$ ($i = 1, 2,...,m$). Then (e.g., Hays, 1988)

$$\Psi = \sum_{i=1}^{m} \frac{(mO[i] - N)^2}{mN} \tag{8}$$

is distributed as chi-square with $mO - 1$ degrees of freedom. Accordingly, if Ψ is larger than some critical value, we reject H_0, and conclude that x is not structureless.

In "*Biconvergent SDPs*", we introduce an important modification to the SDPs used in the sections entitled "*Preliminary tests of structure detection*" and "*Discovering structure in natural images with an SDP*". This new biconvergent SDP makes use of a different statistic than that given by equation (8)—a statistic whose precise form is derived adaptively, in parallel with the basis B.

The tuning procedure

By itself, the structure test detailed above is of little use. It is easy to imagine this test failing to reject a false H_0 due to a poor match between the isometry B and the structures inherent in x. To take an extreme example, if B is the identity on $\mathbb{R}^N$, then $B(G(x)) = G(x)$, in which case the random variables $(B(G(x)))[i]$ are precisely standard Normal; hence, in this case, the test is completely powerless to correctly reject H_0, no matter how false it may be. However, as we shall demonstrate with the examples in the sections entitled "*Preliminary tests of structure detection*" and "*Discovering structure in natural images with an SDP*", it is possible, at least in certain instances, to adaptively tune B, over a series of "learning" trials, to the characteristic structures in a given image population.

The sequence of steps. Prior to tuning, we initialize B_0 to a random, $N \times N$ orthonormal matrix. Then on each learning trial $i = 1, 2,...,$ we

1. *Sample*: randomly select a new image patch $x \epsilon \mathbb{R}^N$ from our target population of images. Then
2. *Test*: use B_i in applying the structure test to x.
3. *Update*: produce the isometry B_{i+1} to be used in the next iteration according to the "update rules," so as to increase the power of the structure test to reject the null hypothesis that x is structureless.

Update rules. There are many possible rules that might be used to transform B_i into B_{i+1}. The rules we have investigated in our simulations are all of the following general form.

For g, a realization of $G(x)$, let $y = B_i g$. Our aim is to increase the deviation from Normality of the histogram of y. Accordingly, we

1. Apply an adaptively evolving transformation Γ to y to produce a "target" vector $y_{target} = \Gamma(y)$, such that
   a. $|y_{target}| = |y|$ (so that y can be rotated to y_{target}),
   b. $|y_{target} - y|$ is "small" (so that B_{i+1} will differ only slightly from B_i), and

c. the histogram of y_{target} deviates more from a standard Normal distribution than does that of y.

Then (for $R_{y,y_{\text{target}}}$ defined by equation (5)) we set

$$B_{i+1} = R_{y,y_{\text{target}}} B_i. \tag{9}$$

This assignment has the desired result that

$$B_{i+1} g = y_{\text{target}} \tag{10}$$

That is, the power of B_{i+1} to detect the structure in x is increased over that of B_i.

Varieties of update transformations. In applying structure detection to natural images, we have experimented with various different update transformations Γ. In sections entitled "*Preliminary tests of structure detection*" and "*Discovering structure in natural images with an SDP*" we shall use transformations of the following form. For any non-negative real number q, define the function f_q: $\mathbb{R} \to \mathbb{R}$ by setting:

$$f_q(r) = \text{sign}(r)|r|^q \tag{11}$$

for all $r \epsilon \mathbb{R}$. Then for any image $v \epsilon \mathbb{R}^N$, set

$$\Gamma_q(v) = \frac{|v|}{|f_q \circ v|} f_q \circ v. \tag{12}$$

The rescaling accomplished by equation (12) insures that $|\Gamma_q(v)| = |v|$, as required in order for $\Gamma_q(v)$ to be reachable from v by a rotation.

If $q > 1$, then Γ_q has the effect of enlarging the magnitude (while preserving the sign) of those pixel values that are greater than 1, and diminishing the magnitude (while preserving sign) of pixel values less than 1. Typically, for $q > 1$, the histogram of $\Gamma_q(v)$ will tend to have higher kurtosis than does the histogram of v.

On the other hand, if $q < 1$, then Γ_q has the opposite effect of diminishing the magnitude (while preserving the sign) of those pixel values that are greater than 1, and increasing the magnitude (while preserving sign) of pixel values less than 1. In this case (for $q < 1$) the histogram of $\Gamma_q(v)$ will tend to have lower kurtosis than does the histogram of v.

In the section "*A simulation using a biconvergent SDP*" we shall examine the performance of 'biconvergent' SDPs whose update procedure differs from that described here. Specifically, these modified SDPs update not only the basis B_i (to produce B_{i+1}), but also a function f_i:$\mathbb{R} \to \mathbb{R}$ (to produce function f_{i+1}) that is both used in Normality testing, and is also used in updating B_i.

Structure detection compared to standard varieties of projection pursuit. Structure detection should perhaps be counted among the data mining techniques called *projection pursuit* methods (e.g., Huber, 1985). There is an assumption, borne out in practice, that linear projections of real-world data sets into arbitrary subspaces tend to have gaussian distributions. Indeed, Diaconis & Freedman (1984) have shown formally that for non-structured data sets, almost all projections are nearly the same and approximately gaussian. Thus, projections that result in non-gaussian distributions often signal, and aid in analyzing the processes that generated the data. Accordingly, in projection pursuit, the aim is to find a subspace, typically (but not necessarily) of low dimensionality, such that the projection into that subspace of the given data set is highly "non-gaussian."

Let $Data \subset \mathbb{R}^N$ be a data set comprising d points, and let D be an $N \times d$ matrix whose column vectors are the points of *Data*. The goal of standard projection pursuit applications is to find an $m \times N$ matrix B (typically, but not necessarily, with $m < N$) such that the matrix $P = BD$ has rows (of length d) whose histograms are highly non-gaussian. The i^{th} row of P is called the *projection* of *Data* onto the i^{th} row vector of B. One chooses a cost function C that is used to evaluate each candidate basis B. C is typically chosen to have low values if the projection matrix BD has rows whose histograms are highly non-gaussian. If one has reason to suppose that there exist projections whose histograms deviate from Normality in a specific way (e.g., due to high kurtosis, or positive skew), then one tailors C to reflect this supposition. Some search procedure is then used to find a basis B for which C is (at least locally) minimized.

The differences between standard projection pursuit procedures and SDPs are illustrated in Fig. 2.

One important difference between structure detection and other variants of projection pursuit is the statistical certification conferred by the gaussian replacement operation. As mentioned above, Diaconis & Freedman (1984) have shown that for IID data sets, almost all projections are nearly the same and approximately gaussian. As concerns the issue of statistical certification, however, the phrase "almost all" is crucial. Consider, for example, a population of IID random images with a highly non-gaussian pixel distribution. For concreteness, imagine that each pixel takes either the value -1 or the value 1 with equal probability. Let x be a random image from this population. Although it is true that for almost all orthonormal bases B, the pixel values of image Bx will be marginally normal in distribution, there obviously exists a highly non-gaussian, full-rank projection of x; in particular, for I the $N \times N$ identity matrix, the pixel values of Ix are all either -1 or 1. In conjunction with observation 1, this observation implies that for bases B other than I, if the pixel values of Bx are marginally normal, then they cannot be independent.

One effect of applying gaussian replacement to x is to remove the trivial solution basis I from the search space. More generally, by substituting the gaussian replacement $G(x)$ for x, we insure that if x is structureless, then for any basis B, the pixel values of $BG(x)$ must by jointly independent, standard Normal random variables. It is precisely this certitude that enables secure statistical inference. It may be of interest to test whether an image population P is truly random (i.e., consists of IID random images). To investigate this issue, one might use an SDP. If, for example, the SDP yields a basis B that succeeds in rejecting H_0 (that the given input image is structureless) at the 0.05 level, for only 7% of a sufficiently large number of images, then one can conclude with high certainty that P is not truly random.

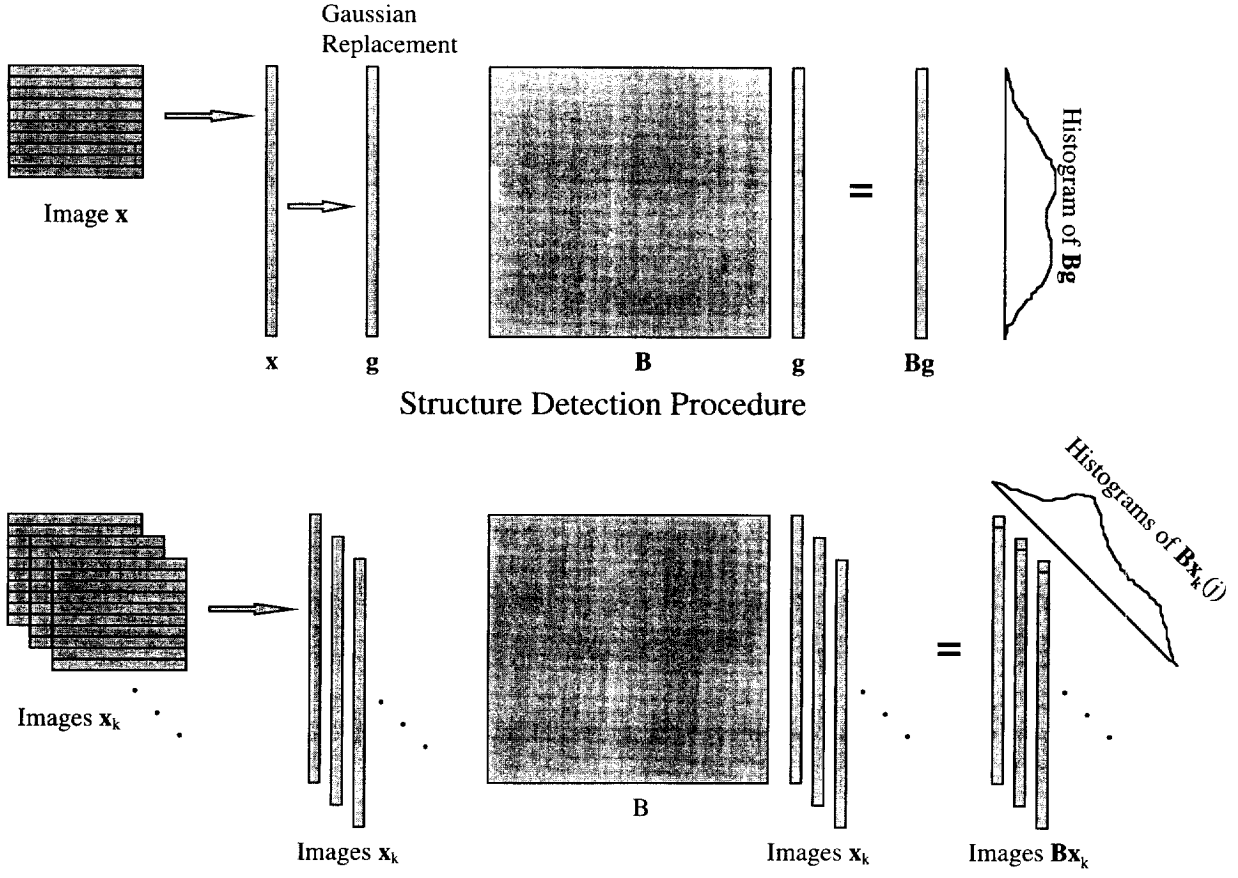


FIGURE 2. The differences between Structure Detection Procedures and standard variants of projection pursuit. In an SDP, an update of the basis B of receptive fields depends on B's response to only a single training image from the target population. By contrast, in most applications of projection pursuit, an update of B depends on B's responses to the entire set of training images, $x_1, x_2, \ldots, x_n$. Standard projection pursuit applications update B so as to increase the deviation from normality of the response histograms of individual receptive fields of B across the entire set of training images x_k. SDPs update B so as to increase the deviation from normality of the histogram of all receptive field responses to the current image. Finally, and crucially, SDPs begin a training trial by transforming the input image x into a realization g of $G(x)$, the gaussian replacement of x. This operation enables statistical assessment of the null hypothesis H_0 that x is devoid of structure.

Another important difference is that the SDP uses a search procedure that is computationally very efficient but unlikely to converge to as nearly optimal a solution as other sorts of projection pursuit. In the current SDP implementation, on each training trial i, the basis B_i is transformed into a slightly different basis B_{i+1} that does a little better than did B_i at rejecting the null hypothesis H_0 that the current input x_i is devoid of structure. This update is made without reference to any of the previous inputs x_j, $j < i$. Thus it is entirely possible that B_{i+1} might perform worse than B_i at rejecting H_0 for some or all of these previous x_j. Clearly, there is no guarantee that this image-by-image training procedure will converge to a basis B_{final} that is in any sense "optimal" at rejecting H_0.

Most standard projection pursuit applications use more powerful search procedures, procedures that update the basis B reiteratively based on B's responses to the entire ensemble of training images. Thus, at each step, B is applied to all training images; the target cost function is computed for the resulting distribution of responses; then B is modified so as to decrease the value of the cost function. In this way, one attempts to arrive at a basis B that is optimal in the sense that the distribution of B's responses to the ensemble of training images minimizes the target cost function.

Of course, updates in such a procedure are expensive in space and time. A single update requires access to the entire store of training images, and extensive computations are required to modify B.

We have opted for an update procedure that makes no claim to optimality, but that (i) seems likely to yield adequate results in practice; and (ii) operates under the same sorts of time and space constraints that would be likely to limit perceptual organization processes occurring in a real organism. The primary virtue of our update procedure is that it requires minimal computational space and time. Images need not be retained in memory. The only space requirement is memory for the basis B ($O(N^2)$, for N the number of pixels in an image); moreover, the computation used to update the basis B takes only $O(N^2)$

time (quite efficient, when one reflects that to compute the product of two matrices, each the size of B, requires $O(N^3)$ time).

The SDPs used in the sections entitled "*Preliminary tests of structure detection*" and "*Discovering structure in natural images with an SDP*" also differ from other sorts of projection pursuit in their handling of the role usually played by a cost function. In standard applications of projection pursuit, the cost function C plays two roles: first, C is used to evaluate each candidate projection of the entire data set. Second, the emergent projection basis B is modified at each step in the pursuit process so as to decrease C (often, the gradient of C is computed at B, and B is updated along the gradient). In the first SDPs we consider, these two roles are shared by the update transformation Γ, and the p-value resulting from the structure test. It is the p-value from the structure test that is used to evaluate the projection basis B at each step. The structure test gauges the deviation form normality of the histogram of B's responses to the current image. However, the update transformation Γ is always used specifically to increase the kurtosis of B's response distribution. Consequently, if the histogram of B's responses to the current image is of lower kurtosis than the standard normal distribution, then the update performed on B could have the immediate effect of increasing the current p-value.

This awkwardness in behavior is avoided by the biconvergent SDPs to be described in the section "*Biconvergent SDPs*". Biconvergent SDPs simultaneously update the basis B as well as a histogram distortion template f (see "*Preliminary definitions*" for the definition of a histogram distortion template) that is integral both in updating B (the role currently played by the update transformation Γ) and also in assessing the deviation from normality of $B(G(x))$ for each successive training image x.

By promoting high kurtosis response histograms, one might expect to discover a basis B (if it existed) such that the response histogram of any given receptive field in B, taken across all images in the target population, was highly kurtotic. Each receptive field of such a basis B would respond strongly to a few images in the target population, and near 0 to all other images. Such a basis B would provide a code for the target image population that is both sparse and distributed. In practice, however, promoting high kurtosis response histograms tends to discover bases B (if they exist) in which a few receptive fields are chronically activated more highly than the others by images from the target population. In the non-artificial applications we shall describe, the histograms of individual basis elements, taken across samples of images from the target population, tend to be normal, but with different standard deviations. Thus, as will become clear from some of the examples in the following sections, the current implementation of structure detection is likely to isolate a subset of quasi-principal components (the highly active receptive fields) of the target image population.

Preliminary tests of structure detection

Testing the randomness of rand(). As a first test of the structure detection procedure, we applied the method to test the notoriously deficient C programming language, random number generator, rand(). To apply the procedure, we generated images v consisting of 31×31 pixels, each assigned the value 0 or 1. The assignments were made by sampling a sequence of 31 successive integers from rand(). Each integer consists of 31 significant bits. The pixel value $v(i,j)$ was set equal to the j^{th} bit of the i^{th} integer (i.e., $v(i,j) = 1$ if bit j of integer i is 1, and $v(i,j) = 0$ otherwise).

A basis B_{final} was obtained after a training sequence consisting of 20 000 images. The effectiveness of B_{final} was then assessed by applying B_{final} in the structure test to a completely new sequence of 10 000 images. The null hypothesis H_0 that images consisted of jointly independent binary values (i.e., that rand() is truly random) was rejected with a p-value less than 0.08 for 90% of the test images. If rand() were truly random, then the rejection rate should be close to 8%. The obtained rate of rejection (90%) is obviously much higher than might be expected by chance (p infinitesimal). The 15 receptive fields of B_{final} that responded most strongly on average to the test images are shown in Fig. 3.

For example, look at the top left-hand square of Fig. 3. Each of this square's 31 rows contains 31 pixels, which correspond to the 31 significant bits composing an integer drawn from rand(). Successive rows correspond to successive integer draws from rand(). In a given row, bit order increases from left to right. Thus the leftmost pixel of the top row corresponds to the low order bit of the first of 31 successive integers drawn from rand(). Notice that in almost all of these 15 most active receptive fields, the weights corresponding to the low order bit oscillate positive and negative from row to row. This finding reflects a well-known deficiency of rand(): successive draws from rand() alternate strictly between odd and even numbers. More generally, these receptive fields make it clear that the low order five bits of integers returned by rand() are far from jointly independent.

We also applied structure detection to the newer UNIX random number generator, random(), which is reputed to be much better than rand(). Exactly the same procedure was applied. In this case, we were unable to reject the null hypothesis that random() was truly random. (The null hypothesis H_0 that images consisted of jointly independent binary values was rejected with a p-value of 0.08 for 8% of the test images.)

Discovering structures in an artificial image population. As a second test of the procedure, we considered a population of images produced using the binary, orthonormal basis W shown in Fig. 4(a). The i^{th} image in the j^{th} row of Fig. 4(a) ($i,j = 0, 1,..., 15$) is defined for pixels (x,y),$x,y \in 0, 1,..., 15$ by:

$$W_{i,j}[x, y] = w_i[x]w_j[y], \tag{13}$$

for w_0, w_1,..., w_{15} the Walsh basis functions, defined on the set {0, 1,..., 15} (e.g., Gonzalez & Wintz, 1987).

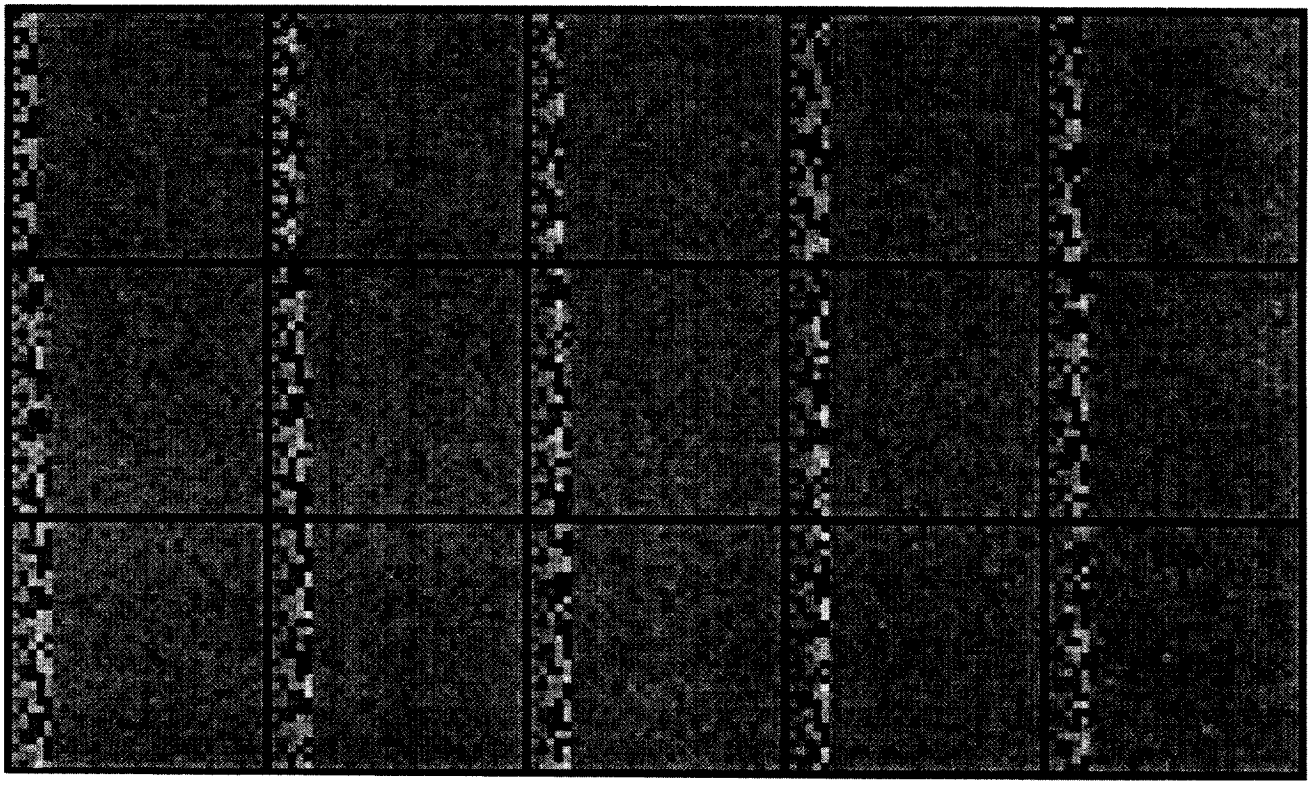


FIGURE 3. The 15 most effective receptive fields derived from an SDP applied to samples of pseudorandom numbers drawn from rand(). Each square contains 31 rows × 31 columns of pixels. The 31 pixels in a given row correspond to the 31 significant bits of an integer drawn from rand(). Successive rows in a square correspond to successive integer draws from rand(). In a given row, bit order increases from left to right. Thus, the leftmost pixel of the top row corresponds to the low order bit of the first of 31 successive integers drawn from rand(). Notice that in almost all of these 15 most active receptive fields, the weights corresponding to the low order bit oscillate positive and negative from row to row. This finding reflects a well-known deficiency of rand(): successive draws from rand() alternate strictly between odd and even numbers. More generally, these receptive fields make it clear that the low order five bits of integers returned by rand() are far from jointly independent.

The images in our test population were constructed by randomly selecting, on each trial, two elements from W, and adding them together with random signs. Specifically, each image v is given by

$$v = \phi w + \rho u \tag{14}$$

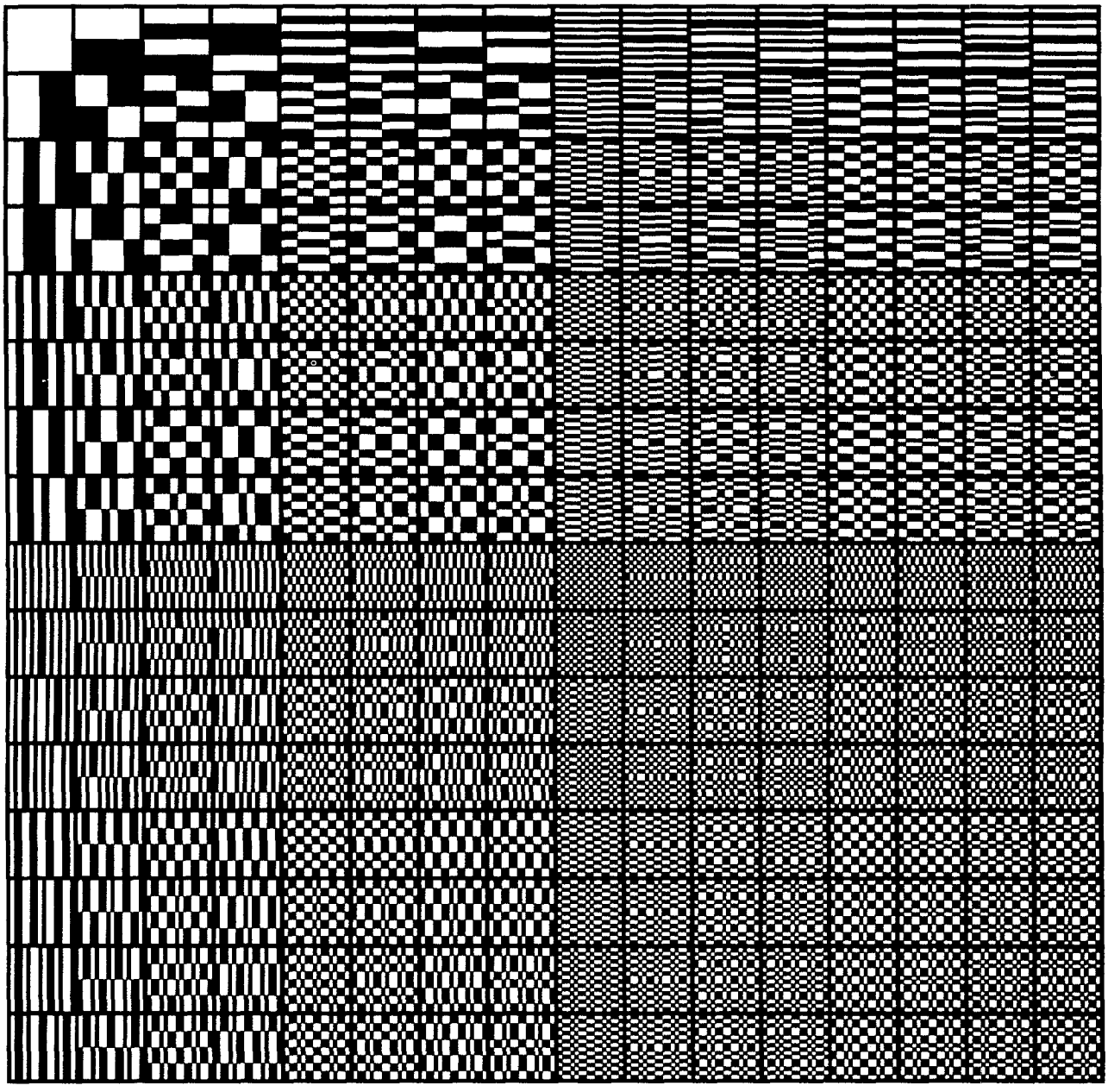
where ϕ and ρ are independent random variables, each assuming the values +1 or −1 with equal probability, and u and w are two random elements of W.

The latent structures in this population are precisely the elements of W. Thus, the basis B_{final} that should be discovered by structure detection is the basis W itself. To elaborate this point, note that for any v given by equation (14), Wv is an image that assigns the value 0 to all but two pixels, each of which is assigned either 1 or −1, with equal probability. Thus, the histogram of Wv is extremely non-gaussian. In particular, Wv's histogram has a huge spike (registering probability 127/128) at 0 (because 254 of 256 of Wv's pixel values are 0), and a spike of size 1/128 at either 1 or −1, or else spikes of size 1/256 at each of 1 and −1. Although the gaussian replacement $G(v)$ differs randomly from v, it is nonetheless to be expected that $WG(v)$ will be highly kurtotic. Indeed, it is clear that W is optimally suited, in the context of the structure test, to reject the null hypothesis that v is structureless.

This simulation provides a useful test of the structure detection procedure. Note that the histogram of each input image in this population has at most three values; thus, for each image, the gaussian replacement procedure introduces dramatic, random intensity changes. One might have supposed that the intensity distortions introduced by gaussian replacement would lead to a value of B_{final} other than W. However, this is not the case. We applied structure detection to a sequence comprising 196,608 images from this population. The resulting basis B_{final} is shown in Fig. 4(b). As is clear, B_{final} converges precisely to W.

DISCOVERING STRUCTURE IN NATURAL IMAGES WITH AN SDP

In this section we shall describe the results of applying an SDP to a collection of natural images. We begin by describing the image selection and preparation procedure. Next, we describe details of the specific computation used. Finally, we present the results of the simulation and interpretation.



(a)

FIGURE 4(a)—*Legend opposite.*

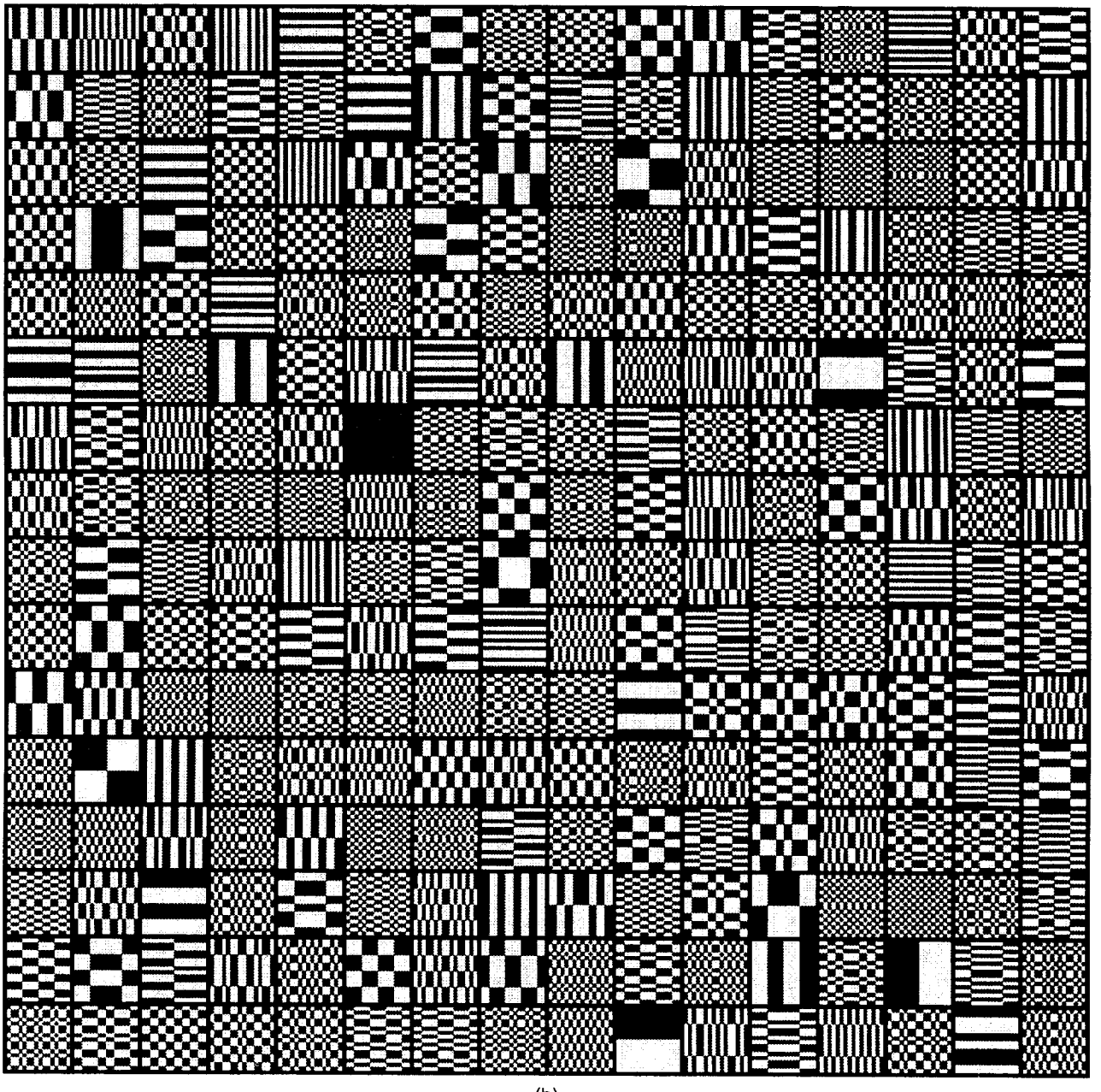
Procedures

Image selection and preparation. The images we use are drawn from the PhotoDisc "Starter Kit" CD-ROM (PhotoDisc, 1995), which contains roughly 9000 digitized photographs of diverse scenes. We use the following criterion to select our images. A given image is included in our ensemble if, and only if, it contains no man-made objects or man-produced patterns. In particular, we exclude any images containing clothing or paintings. This procedure yields a set of 958 images comprising landscapes, plants, animals and portions of the human body. The images varied in size: most were around 400×400 pixels.

Each of these colored photographs is first converted into an eight-bit, digitized, grayscale image. Each grayscale image is then parceled into a collection of 16×16 pixel patches, yielding a total of $8 \times 20 \times 958 = 153\,280$ patches. Patches were then placed into a quasi-random sequence by performing the following operation eight times:

(i) We select a random subset of 20 patches from each of our 958 images, making sure that none of these 20 patches have ever been included previously in the sequence. Then we

(ii) Randomly order the resulting set of 19 160 patches; and



(b)

FIGURE 4(b)

FIGURE 4. Applying an SDP to artificial images. (a) An orthonormal basis W of binary images. An SDP was applied to an artificial population of images in which random pairs of elements of W were added together with weights randomly equal to 1 or -1. The source structures in this population are precisely the elements of W. Thus, the basis B_{final} that should be discovered by the SDP is the basis W itself. (b) The basis B_{final} obtained. As is clear, B_{final} converges accurately to W.

(iii) Append the new subsequence to the previously generated sequence.

Finally, all perfectly uniform patches (all patches whose pixels are all assigned a single value) were removed from the sequence. Rationale: any uniform image is structureless in any image population. This point is underscored by noting that the gaussian replacement of a uniform patch is literally a standard normal IID image. Thus, for such patches, the null hypothesis is necessarily true. This screening procedure removed 0.5% of all image patches.

The remaining sequence contained a total of 152 508 16×16 pixel patches. The training sequence comprised the first 133 443 of these patches. The testing sequence comprised the remaining 19 065 patches.

Training. In the current application of structure detection, we initialize B_0 to a 256×256 random, orthonormal matrix. Each row of B_0 corresponds to a single 16×16 pixel receptive field, which will be kept

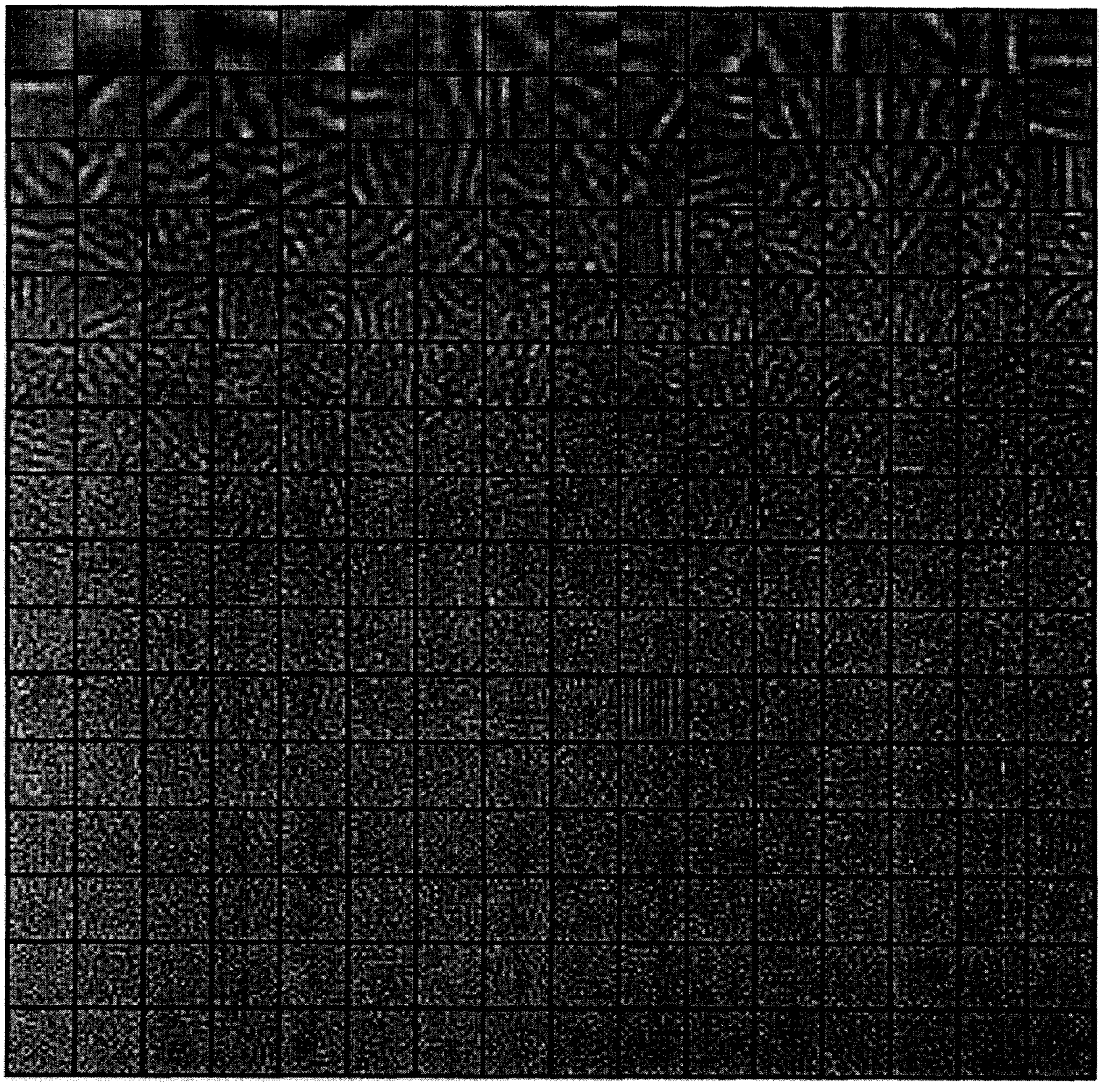


FIGURE 5. Applying an SDP to natural images. The receptive fields of basis B_{final}, derived from applying an SDP to a sequence of randomly chosen, 16×16 pixel patches of natural image. These receptive fields are ordered, from left to right within a row, and from top to bottom, in terms of the average absolute value of their responses to the test images. The most active receptive fields are selective for oriented, low and moderate spatial frequencies, reflecting the prevalence of low spatial frequencies in natural images. Those receptive fields showing high activation levels display orientation selectivity and spatial localization similar to simple cell receptive fields. Only the 40 most active receptive fields had an average activation level greater than 1; the other (less active) receptive fields issued responses consistently very near 0. Thus, although the code for natural images provided by these receptive fields is sparse, it is not distributed.

orthonormal with respect to all the other rows as B_i evolves throughout the procedure.

On the i^{th} training trial, we follow the usual sequence of steps described in "*The tuning procedure*". Specifically, we

1. *Sample* by reading in the i^{th} patch x in the training sequence,
2. *Test* by applying the structure test to the gaussian replacement of x; and
3. *Update* B_i using an update transformation Γ_q (see "*The tuning procedure*") that is altered gradually during the training process. The parameter q is restricted to the range [1, 1.25] on any given training trial. As the basis B_i is transformed into an increasingly powerful structure-detector, the value of q is adaptively reduced. This tends to yield smaller and smaller changes to B_i as training proceeds. We have no reason to believe that the particular mechanism of reducing q influences the outcome of the simulation in any major way.

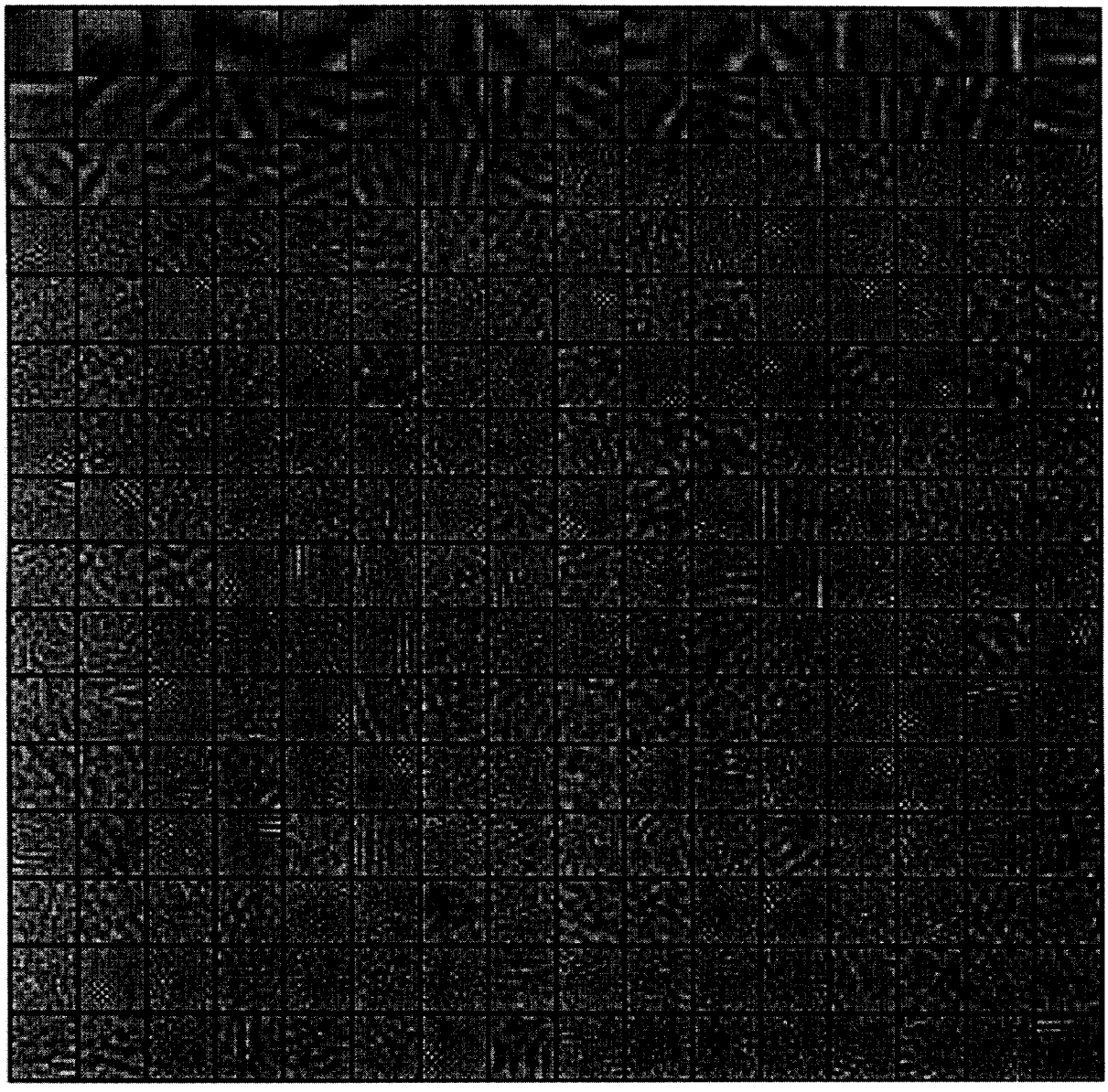


FIGURE 6. Extraction of residual, high-frequency structure from the training images. Although most of the energy in the training images is concentrated in the 40 most active receptive fields obtained in Fig. 5, there remains high spatial-frequency structure in the residual images. This high-frequency structure can be detected by reiterating the SDP on the set of residual images. The result of this operation is shown here. The first 40 receptive fields in this figure are identical to the first 40 in Fig. 5. The other receptive fields result from applying an SDP to the sequence of residual training images. These new receptive fields are more obviously structured than their counterparts in Fig. 1. This new, high-frequency ensemble enables us to reject H_0 with a p-value of 0.05 for 37% of the test images. Although this performance level is much lower than was obtained for the original training set, it is still highly significant.

Assessment. The entire sequence of training trials yielded a matrix B_{final} which was supposed to capture the characteristic structures inherent in the images of the training sequence. To assess B_{final}'s ability to reject H_0, we proceeded to use B_{final} in applying the structure test to each patch in the testing sequence (none of which had been presented during training).

Results

Overall performance. H_0 was rejected by B_{final} in the structure test at the 0.005 level of significance, for 87.9% of the patches in the testing sequence. The average p-value over all patches in the testing sequence was 0.024, indicating that B_{final} does a creditable job at capturing the structure in the image population.

Low-spatial-frequency receptive fields. The receptive fields of B_{final} are shown in Fig. 5. These receptive fields are ordered in terms of the average magnitude of their responses to the training images. As is evident, the most active receptive fields are selective for oriented, low and moderate spatial frequencies. This reflects the well-known prevalence in natural images of low spatial

frequencies (e.g., Field, 1987). Those receptive fields showing high activation levels are similar in structure to simple cell receptive fields (i.e., they tend to be oriented and spatially localized). Interestingly, only the 40 most active receptive fields had an average activation level (across the training image set) greater than 1 (the value expected for all receptive fields if H_0 were true). Indeed, most of the other (less active) receptive fields issued responses consistently very near zero. Note that these less active receptive fields are devoid of any obvious structural components; they show little orientation tuning and little spatial localization. Evidently, the typically high kurtosis of the histograms obtained by applying B_{final} to the testing images results from having a handful of receptive fields issuing high responses to most images, while all other receptive fields issue responses consistently near 0. Thus, although the code for natural images provided by these receptive fields is sparse (only a few receptive fields are strongly activated by any given input image), it is not distributed (most units are never very active).

The fact that few receptive fields (only 40) were, on average, more active than was to be expected by chance suggests that

1. The training images were generated primarily by only a few latent structures (corresponding to the active receptive fields), which were successfully discovered by the structure detection procedure.
2. The resulting code is fairly compact, because only the responses of this small set of active receptive fields are required to approximately code the original images.

High-spatial-frequency receptive fields. Although most of the energy in our training images is concentrated in the 40 most active receptive fields obtained, there remains high spatial frequency structure in the residual images (i.e., in those components of the original images that are not captured by the 40 most active receptive fields). Even though high frequencies contribute scant overall energy to natural images, the information carried by high frequencies can be of crucial importance to an observer.

High-frequency structure can be detected by reiterating the structure detection procedure on the set of residual images. Specifically, let B_{residual} be the 216×256 matrix comprising all those receptive fields (rows) of B_{final} whose average response energy to the testing sequence was less than 1.0. We project each input image of our training sequence into the space spanned by the rows of B_{residual}. Then we apply structure detection to the resulting sequence of residual images, starting with matrix $B_0 = B_{\text{residual}}$. Thereby, all subsequent, updated B_i's remain orthonormal to the 40 elements of B_{final} excluded from B_{residual}.

Although the residual images are much lower in energy than the original images, they remain rich in delicate structure, structure that could not be detected previously because of the predominance of the low-frequency response energy. In essence, this operation removes from each image in our training and testing sequences those components whose activation levels were consistently determining the tails of our response histogram.

The result of this reapplication of structure detection is shown in Fig. 6 (compare with Fig. 5). The first 40 receptive fields in Fig. 6 are reproduced from Fig. 5 to facilitate the comparison of the 216 remaining receptive fields in each of the two figures. The 216 receptive fields that result from reiterating the structure detection procedure on the sequence of residual training images are more obviously structured than their counterparts in Fig. 5. The 216 new receptive fields enable us to reject H_0 (the null hypothesis that the residual images are devoid of structure) with a p-value of 0.05 for 37% of the images in the testing sequence. Although this performance level is much lower than was obtained for the original training set, it is still highly significant.

A SEARCH PROCEDURE WITHOUT GAUSSIAN REPLACEMENT

Procedure

A novel aspect of structure detection is the gaussian replacement operation. It is this processing step that enables one to submit the current input to a statistical test for the presence of structure. Here we perform the training procedure for natural images without applying gaussian replacement. We use precisely the same training procedure and precisely the same training set of images as in the previously described training procedure in which gaussian replacement was applied.

The result of omitting gaussian replacement in the training procedure is shown in Fig. 7. The resulting receptive fields have been ordered in terms of their average response energy to the test images. The new basis is very similar to the basis obtained using gaussian replacement.

Results

That the new receptive fields are at least as sensitive to the characteristic structures in natural images is indicated by a simulation in which this new ensemble of receptive fields is applied in the structure test to the test images. Thus, although gaussian replacement was not used in the training procedure that yielded the receptive fields B shown in Fig. 7, we can nonetheless use gaussian replacement in the context of the structure test to test the sensitivity of B to characteristic structures in the population of natural images. The total number of test patches was 19 160, of which 92 were rejected because they were uniform in intensity, leaving a total of 19 068 patches to which the structure test was actually applied. The average p-value across these images was 0.023, and the null hypothesis that input pixels were jointly independent, identically distributed random variables was rejected at the 0.005 level of significance for 89.26% of the images.

Evaluation of the results

The results without gaussian replacement are at least as good as were previously achieved with the basis obtained with the training procedure that used gaussian replacement. This suggests that gaussian replacement during training is not crucial to the detection of structure. However, to test the effectiveness of the resulting receptive fields, gaussian replacement is required in the context of the structure test. It is worth noting, however, that irrespective of how one obtains a given orthonormal basis B, the structure test (which crucially involves gaussian replacement) provides a useful new tool for assessing the sensitivity of B to the structures in a given image population P. One can use any candidate orthonormal basis B in applying the structure test to a random set of test images from P. B's sensitivity to the characteristic structures of P is gauged by B's overall effectiveness at rejecting H_0 in the structure test.

Finally, we submit that people are extremely sensitive, not just to characteristic structures in the environment, but to the *absence* of such structure when some generative aspect of the visual input is random. For instance, we easily sense that the precise locations of leaves on a bush are random. For purposes of responding adequately to stimuli, it is critical to be able to discriminate such random aspects of the visible world from those that are systematic. Gaussian replacement, or some analogous strategy for statistical certification, would be very useful for making such judgments.

BICONVERGENT SDPS

The SDPs used in the preceding sections have several interrelated weaknesses. First, the structure test is not used to guide the development of the basis B; it is used only to assess the efficacy of B at rejecting H_0 for the current image x. To steer the evolution of B, each of the previous SDPs has made use of an update transformation Γ_q ("*The tuning procedure*") that was designed to drive the histogram of receptive field outputs away from normality by increasing its kurtosis. Although the results obtained in our previous simulations validate this strategy (insofar as they have yielded Bs that were able to reject H_0 with a success rate greater than chance), the strategy is nonetheless *ad hoc*. There certainly is no guarantee that the optimal way to drive a response histogram away from normality is by increasing its kurtosis.

It is useful to review the precise procedure that was used to generate receptive field histograms (of ever increasing kurtosis) before considering procedures that generate optimum histograms (for rejecting H_0). Consider a new image x. Let g be a particular realization of random gaussian replacement $G(x)$, and let b be a row vector of the basis of receptive fields B; i.e., b is a particular receptive field. Previously, B was updated to drive the receptive field response (a real number) $b \cdot g$ away from zero if $|b \cdot g| > 1$ and towards zero otherwise.

The biconvergent SDPs described in this section impose no a priori assumptions about receptive field response histograms. Biconvergent SDPs simultaneously discover a basis B concurrently with an associated histogram distortion template f, such that B and f together are used to reject H_0 in a histogram-specific structure test. No a priori assumption is made about how receptive field responses should be altered to achieve a non-gaussian receptive field response histogram.

The basis B and the histogram distortion template f evolve in tandem, each influencing the refinement course of the other. The emerging function f embodies information about the ways in which the histogram of receptive field outputs Bg deviates from normality; this f-embodied information is essential in guiding the evolution of B. In turn, the histograms of the receptive field outputs Bg are essential in refining f. The two structures B and f thus crystallize co-dependently.

Throughout this section, we assume an image population P from which images x, each comprising N pixels, are drawn at random. $G(x)$ continues to denote the gaussian replacement of image x, and g will denote a specific realization of the random image $G(x)$. As usual, B denotes an $N \times N$ orthonormal basis subject to modification by an SDP. H_0 denotes the null hypothesis that the current image x is structureless.

The structure test used by a biconvergent SDP

To evaluate the gaussianness of a histogram and to create deviations of a given histogram from a gaussian histogram, it is convenient to operate in an interval $(-1, 1)$ in which the expected histogram of a gaussian distribution is flat. To move the histogram problem into this space, we note that for any random variable Y with continuously increasing distribution function F, the random variable $F(Y)$ is uniformly distributed on $(0, 1)$. It is much more convenient to operate on the random variable $U = 2F(Y) - 1$ which is uniformly distributed on $(-1, 1)$ than on Y itself. As previously noted, under H_0, the receptive field outputs $y = Bg$ are standard normal IID. In this case, for Φ, the standard normal distribution function, the pointwise transformed receptive field outputs $u = 2\Phi \circ y - 1$ consist of jointly independent random variables, all uniformly distributed on $(-1, 1)$. Thus, under H_0, the expected histogram of u is flat across the interval $(-1, 1)$.

As above, let g be a realization of $G(x)$ for an input image x. Let y be the vector of receptive field responses:

$$y = Bg \tag{15}$$

Correspondingly, u indicates the vector of pointwise transformed receptive field outputs:

$$u = 2\Phi \circ y - 1. \tag{16}$$

We shall sometimes use the symbol u without explicitly introducing the corresponding input image x and vector y of receptive field outputs. The important thing to remember is that under H_0, y is standard normal IID, and u is also IID, uniformly distributed on $(-1, 1)$.

Let f be a histogram distortion template (see "*Preliminary definitions*"). It is immediately apparent from Eq. (2) that, for any random variable U uniformly

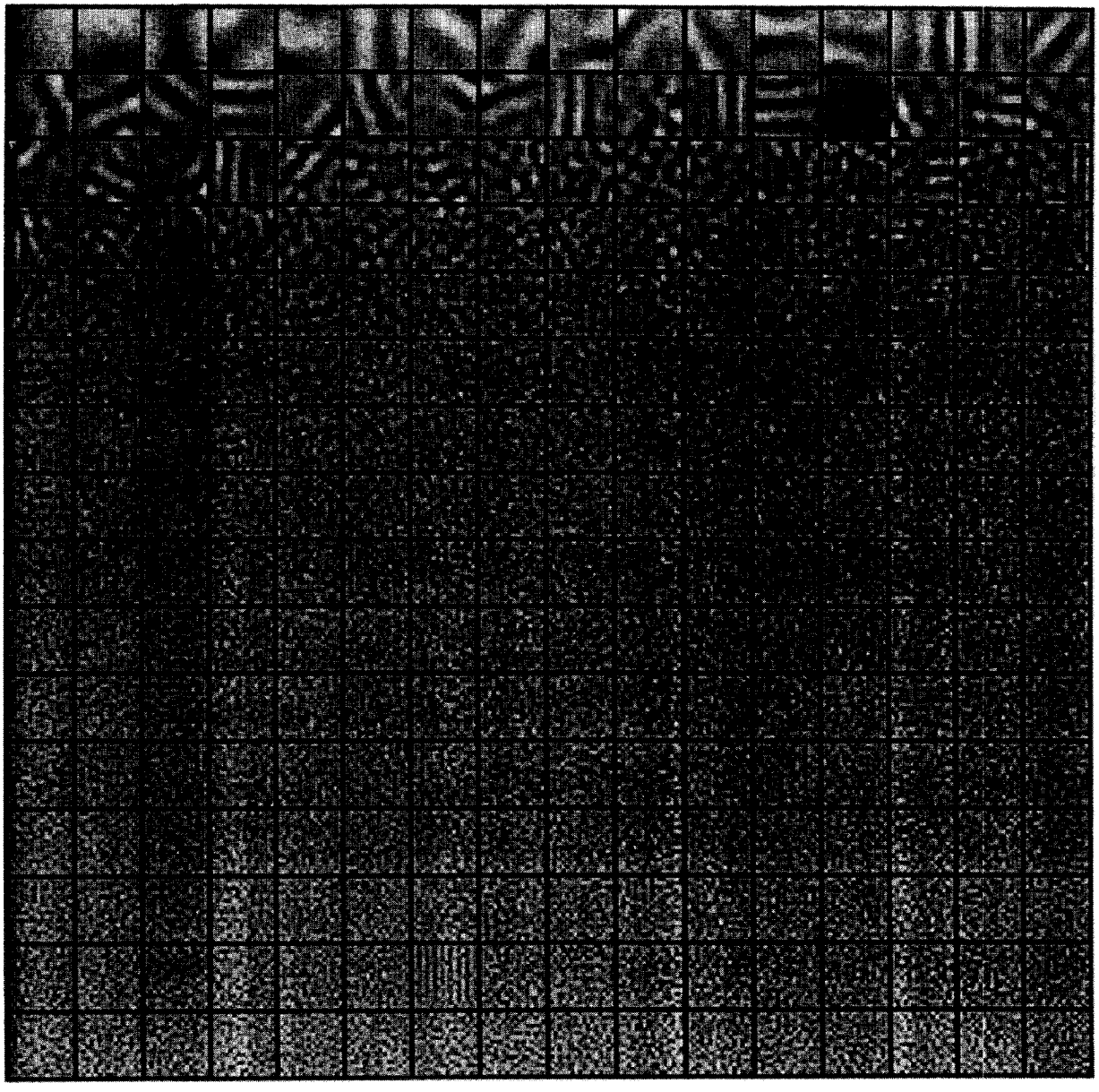


FIGURE 7. Applying an SDP to natural images without gaussian replacement. The receptive fields of basis B_{final}, derived from applying an SDP to a sequence of randomly chosen, 16×16 pixel patches of natural image without gaussian replacement.

distributed on $(-1, 1)$, $E[f(U)] = 0$, and $var[f(U)] = 1$. Under H_0, the transformed receptive field outputs $f \circ u$ consist of jointly independent, identically distributed random variables all with mean 0 and standard deviation 1. When the number of pixels N in an image is moderately large, the Central Limit Theorem implies that the statistic:

$$\Delta_f(u) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} (f \circ u)[j] \tag{17}$$

will have, approximately, a standard normal distribution, where j is an index over receptive field outputs. If H_0 is false, and the vector u is not actually uniform IID, but rather tends to concentrate values primarily in regions of $(-1, 1)$ where f is positive, then $\Delta_f(u)$ will tend to assume large values. For this reason, we call $\Delta_f(u)$ the f-distortion of u.

The biconvergent SDP seeks to derive a basis B with an associated histogram distortion template f, such that the f-distortion of u is consistently, improbably high across different input images x. The structure test rejects H_0 for a given image x if the p-value, $p = 1 - \Phi(\Delta_f(u))$ is less than some critical value (we arbitrarily adopt a critical value of 0.005 in reporting simulation results). In other words, the structure test applies a positive, one-tailed z-test to $\Delta_f(u)$ to test H_0 for input image x.

The coding of histogram measures

Before we discuss the search procedure used by biconvergent SDPs, we should mention that the histogram distortion templates used in the SDPs to be

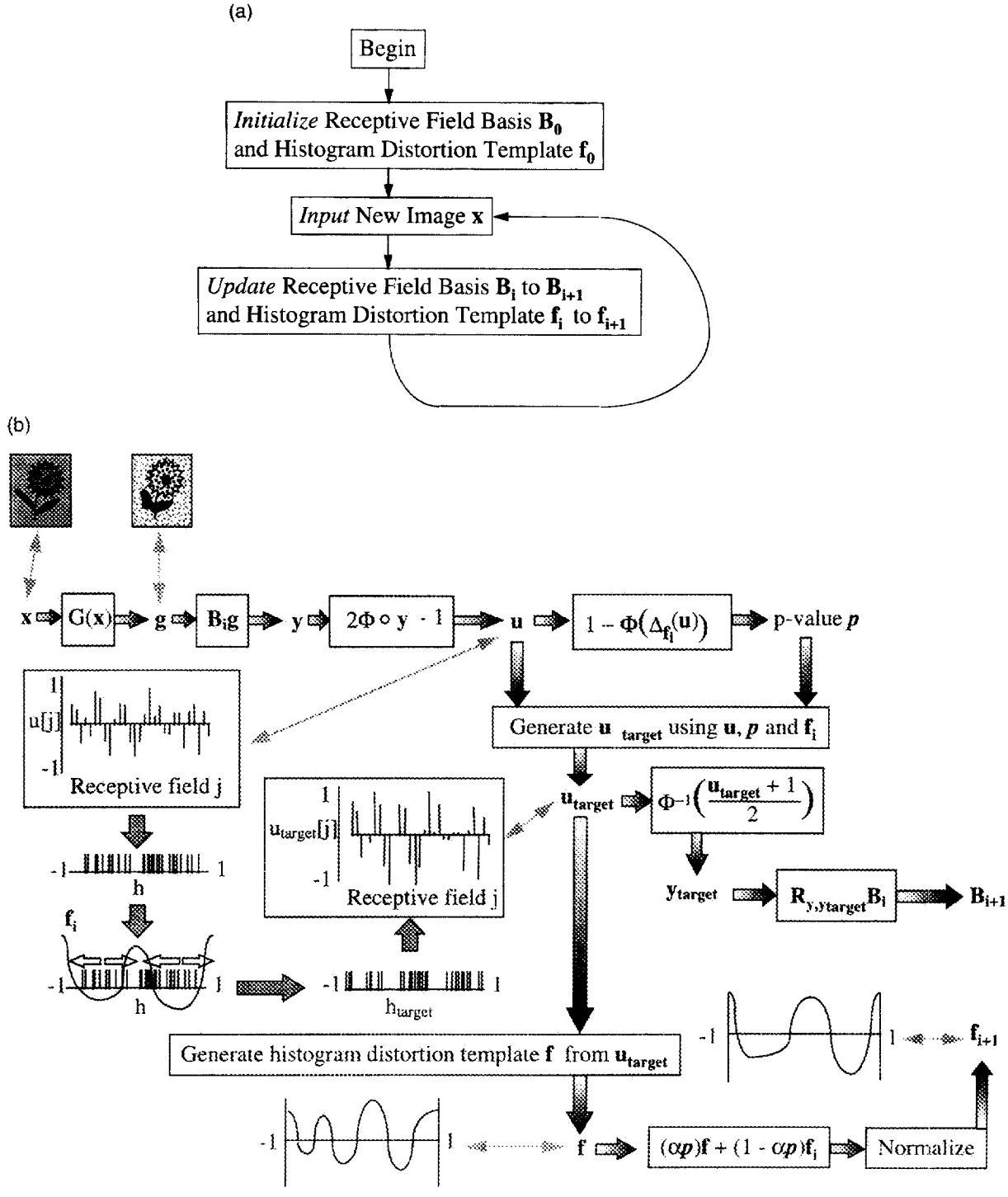


FIGURE 8. The flow of processing in a biconvergent SDP. (a) Global flow of control in a biconvergent SDP. (b) The processing that occurs in updating B_i to B_{i+1} and f_i to f_{i+1}. Input image x is transformed into a realization g of $G(x)$, the gaussian replacement of x. Receptive field basis B_i is applied to g to produce the vector y of receptive field responses. Under H_0, the vector $u = 2\Phi \circ y - 1$ contains jointly independent coordinate values, all uniformly distributed on the interval $(-1, 1)$. Again, under H_0, the f_i-distortion of u, $\Delta_{f_i}(u)$, is a standard normal random variable. Thus, $1 - \Phi(\Delta_{f_i}(u))$ is the p-value resulting from a positive one-tailed z-test applied to the f_i-distortion of u. We proceed to generate a vector u_{target} in the neighborhood of u, but with greater f_i-distortion than u. f_{i+1} is produced by: (i) generating a histogram distortion template f tuned specifically to the various concentrations of values in h_{target}, the histogram of u_{target}; and then (ii) taking a weighted sum of f with f_i, where the sum is heavily dominated by the prior histogram distortion template f_i. To produce the new receptive field basis B_{i+1}, we first transform u_{target} into $y_{\text{target}} = \Phi^{-1}\left(\frac{u_{\text{target}}+1}{2}\right)$. y_{target} Is a vector of receptive field outputs similar to y, but such that the f_i-distortion of $2\Phi \circ y_{\text{target}} - 1$ is greater than that of $2\Phi \circ y - 1$. Then we set $B_{i+1} = R_{y,y_{\text{target}}} B_i$. This has the desired result that $B_{i+1}g = y_{\text{target}}$, making B_{i-1} more effective at rejecting H_0 for x than was B_i.

described are linear combinations of the Legendre polynomials of orders 1 through 12, which we shall denote λ_1, λ_2,...,λ_{12}. The Legendre polynomials are chosen for convenience; other coding schemes would doubtless serve as well. These functions are orthonormal on the interval $(-1, 1)$. That is, for $j, k \in \{1, 2,...,12\}$,

$$\int_{-1}^{1} \lambda_j(r)\lambda_k(r)dr = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases} \quad (18)$$

Any histogram distortion template

$$f = \sum_{k=1}^{12} w_k \lambda_k \quad (19)$$

satisfies Eq. 2(a) by dint of the fact that Eq. 2(a) is satisfied by each of the Legendre polynomials λ_k, $k = 1$, 2,..., 12. To insure that f satisfies Eq. 2(b), we must have $|w| = 1$, for $w = (w_1, w_2,..., w_{12})$.

In the biconvergent SDPs to be described, histogram distortion templates are coded by their Legendre polynomial coefficients. For example, the histogram distortion template f of equation (19) would be coded by the vector w. We call w the Legendre code of f.

The search procedure used by biconvergent SDPs

The search procedure used by a biconvergent SDP is diagrammed in Fig. 8. The overall flow of control is given in Fig. 8(a). As this figure indicates, images x are read in, one at a time, and are used to update both the current basis of receptive fields, and also the current histogram distortion template. The complexities of the process are all embedded in the Update box of Fig. 8(a). The details of the processing that occurs in this Update box are presented in Fig. 8(b). In Fig. 8(b) B_i and f_i denote the basis of receptive fields and associated histogram distortion template that resulted from iteration $i - 1$ of the SDP.

Initialization. At the start of training, B_0 is initialized to a random, $N \times N$, orthonormal basis, and the Legendre code of f_0 is initialized to a random, normalized vector of length 12.

The update procedure (Fig. 8b). Each training trial proceeds as follows. Input image x is first transformed into the image g, a realization of the gaussian replacement $G(x)$. The basis B_i is then applied to g, yielding a vector y of receptive field outputs. (Under H_0, y is standard normal IID.) Vector y is converted into vector u by applying the pointwise transformation $u = 2\Phi \circ y - 1$. (Under H_0, u is IID, with each component uniformly distributed on $(-1, 1)$.) Next, we apply the structure test, computing the f_i – distortion of u, $\Delta_{f_i}(u)$, and the associated p-value, $p = 1\text{-}\Phi(\Delta_{f_i}(u))$.

These computations can be regarded as preliminary to the production of B_{i+1} and f_{i+1}. This process, which begins directly after the structure test, involves the following main steps:

1. An *update vector of transformed receptive field outputs* u_{target} (with pixel values confined to the interval $(-1, 1)$) is selected in the neighborhood of u so that the f_i-distortion of u_{target} is greater than the f_i-distortion of u:
$$\Delta_{f_i}(u_{\text{target}}) > \Delta_{f_i}(u). \quad (20)$$
Thus the target p-value $p_{\text{target}} = 1 - \Phi(\Delta_{f_i}(u_{\text{target}}))$ is less than p.
2. Then, as in the previous SDPs, B_i is updated so that applying B_{i+1} to g results in a p-value equal to p_{target} (which is lower than the p-value p obtained by applying B_i to g). Specifically, applying B_{i+1} to g yields a vector y_{target} of receptive field outputs (instead of the receptive field outputs y produced by applying B_i to g) such that $2\Phi \circ y_{\text{target}} - 1 = u_{\text{target}}$, whereas previously $2\Phi \circ y - 1 = u$.
3. The updated histogram distortion template f_{i+1} is produced by taking a weighted average of the previous distortion template, f_i, with an estimate of the histogram distortion template that would optimally reject H_0 for the updated, transformed receptive field outputs u_{target}.

Choosing the updated, transformed output vector u_{target}. Let f_i' denote the derivative of f_i. To obtain the u_{target}, for each receptive field j, we first set

$$\hat{u}_{\text{target}}[j] = \begin{cases} u[j] + \text{sign}\,(f_i'(u[j]))\alpha p & \text{if } -1 < \text{this value} < 1, \\ 0.9999999999 & \text{if the above value} \geq 1, \\ -0.9999999999 & \text{otherwise,} \end{cases} \quad (21)$$

where p is the p-value obtained from the structure test applied to x at the beginning of this training trial, and α is a parameter governing the size of the adjustments made to B_i. At the start of training α is relatively large (0.05 in our simulations); whereas, by the end of training, when the refinement of B_i is nearly complete, α is small (0.0005 in our simulations). All adjustments to α are scheduled prior to the beginning of training.

The image $\hat{u}_{\text{target}}$ has the desired property that $\Delta_{f_i}(\hat{u}_{\text{target}}) > \Delta_{f_i}(u)$. As the corresponding vector $\hat{y}_{\text{target}}$ (given by equation (22)) may not have the same length as y, it is normalized. Toward this end, we set

$$\hat{y}_{\text{target}} = \Phi^{-1} \circ \left[\frac{\hat{u}_{\text{target}} + 1}{2}\right], \quad (22)$$

and then set

$$y_{\text{target}} = \frac{|y|}{|\hat{y}_{\text{target}}|}\hat{y}_{\text{target}}. \quad (23)$$

The norm adjustment performed in equation (23) insures that y_{target} can be reached by a rotation from y (the original vector of receptive field outputs). We now set

$$u_{\text{target}} = 2\Phi \circ y_{\text{target}} - 1. \quad (24)$$

There are a few things to note about equation (21). First, almost all values $\hat{u}_{\text{target}}[j]$ are shifted by exactly the same distance from their respective starting values $u[j]$. The only exceptions are those coordinate values that end up being either 0.9999999999 or −0.9999999999 because their corresponding u coordinate values were

very close to either 1 or -1. The direction of the shift follows the sign of f_i's derivative. The histogram distortion templates that can be achieved as linear combinations of the Legendre polynomials of order 1 through 12 all have relatively slowly changing derivatives (attaining at most 13 extrema over the interval $(-1, 1)$). Thus, if α is relatively small, we have

$$f_i(\hat{u}_{\text{target}}[j]) > f_i(u[j]) \tag{25}$$

for almost all pixels j. The only exceptions will be those pixels j for which the value $u[j]$ happens to be so close to a maximum of f_i that $u_{\text{target}}[j]$ ends up on the opposite side of that maximum, lower in value than $u[j]$. The update image u_{target} will be nearly identical to u_{target}, conferring high probability to the event that the f_i-distortion of u_{target} will be greater than the f_i-distortion of u:

$$\Delta_{f_i}(u_{\text{target}}) = \sum_{j=1}^{N} (f_i \circ u_{\text{target}})[j] > \sum_{j=1}^{N} (f_i \circ u)[j] = \Delta_{f_i}(u). \tag{26}$$

Updating B_i. In the course of producing update image u_{target}, we also produce target receptive field output vector y_{target} [given by equation (23)]. We now set

$$B_{i+1} = R_{y,y_{\text{target}}} B_i. \tag{27}$$

[See equation (5) for the definition of $R_{y,y_{\text{target}}}$.] This assignment has the desired effect that $B_{i+1}g = y_{\text{target}}$, and hence that $2\Phi \circ y_{\text{target}} - 1 = u_{\text{target}}$, leading to the result that the p-value produced by applying the structure test to image x, using basis B_{i+1} with histogram distortion template f_i will almost certainly be smaller than the p-value obtained using basis B_i with histogram distortion template f_i.

Updating f_i. Following convention, let δ denote the delta function: i.e., $\delta(0)$ is an impulse of infinitesimal width and unit area (hence infinite height), whereas $\delta(r) = 0$ for all $r \neq 0$. Thus, for any function $f:\mathbb{R} \to \mathbb{R}$,

$$\int_{-\infty}^{\infty} \delta(r) f(r) dr = f(0) \tag{28}$$

Let h_{target} denote the irregular comb function given for any $r \in (-1, 1)$ by

$$h_{\text{target}}(r) = \sum_{j=1}^{N} \delta(r - u_{\text{target}}[j]). \tag{29}$$

Think of h_{target} as the histogram of u_{target} (with bins infinitesimal in width, and all bins with frequency 0, except those located precisely at values $u_{\text{target}}[j]$, which have frequency 1).

We next compute the projection of h_{target} into our space of candidate histogram distortion templates. Specifically, we set

$$f = \sum_{k=1}^{12} w_k \lambda_k, \tag{30}$$

where the Legendre code $w = (w_1, w_2, \ldots, w_{12})$ of f is obtained by normalizing the vector $\hat{w} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{12})$, whose coordinates are given by

$$\hat{w}_k = h_{\text{target}} \cdot \lambda_k = \int_{-1}^{1} \lambda_k(r) h_{\text{target}}(r) dr = \sum_{j=1}^{N} \lambda_k(u_{\text{target}}[j]). \tag{31}$$

Thus, f is the histogram distortion template (available in the space of candidate templates) that correlates most strongly with the update image histogram, h_{target}. Consequently, f will tend to have maxima at points in $(-1, 1)$ where h_{target} has high concentrations of spikes, and minima where h_{target} has low spike concentrations.

Of course, we want f_{i+1} to be influenced primarily by f_i, and only mildly affected by the current distortion template, f. Accordingly, we set

$$f_{i+1} = (\alpha p) f + (1 - \alpha p) f_i, \tag{32}$$

where, as in equation (21), α is a parameter that ranges, according to a preset schedule, from 0.05 down to 0.0005 over the course of training, and p is the p-value derived from the structure test at the beginning of the current training trial.

Finally, note that the order in which B_{i+1} and f_{i+1} are produced makes no difference. Each is generated independently from u_{target}.

A SIMULATION USING A BICONVERGENT SDP

A biconvergent SDP using the update procedure described in the previous section was applied to the same body of natural images as was used in the simulation of the section entitled "*Discovering structure in natural images with an SDP*".

Image selection

In this simulation, we used essentially the same image set as was described in "*Discovering structure in natural images with an SDP*". The only difference was as follows: in the previous simulation, 0.5% of the images were removed from the testing and training sequences because they were uniform in intensity (all 16×16 pixels assumed the same intensity), and hence literally devoid of structure. These patches were left in the set for the current simulation. This places the biconvergent SDP at a disadvantage compared with the previous SDP; however, as will become evident, the procedure is quite robust with respect to the inclusion of such structureless images.

Assessment

The sequence of training trials yielded:

1. The matrix B_{final} whose receptive fields (basis elements) are shown in Fig. 9.
2. The histogram distortion template shown in Fig. 10.

To assess the effectiveness of B_{final} in conjunction with f_{final} at rejecting H_0, we proceeded to use B_{final} with f_{final} in applying the structure test to each patch in the testing sequence (none of which had been presented during training).

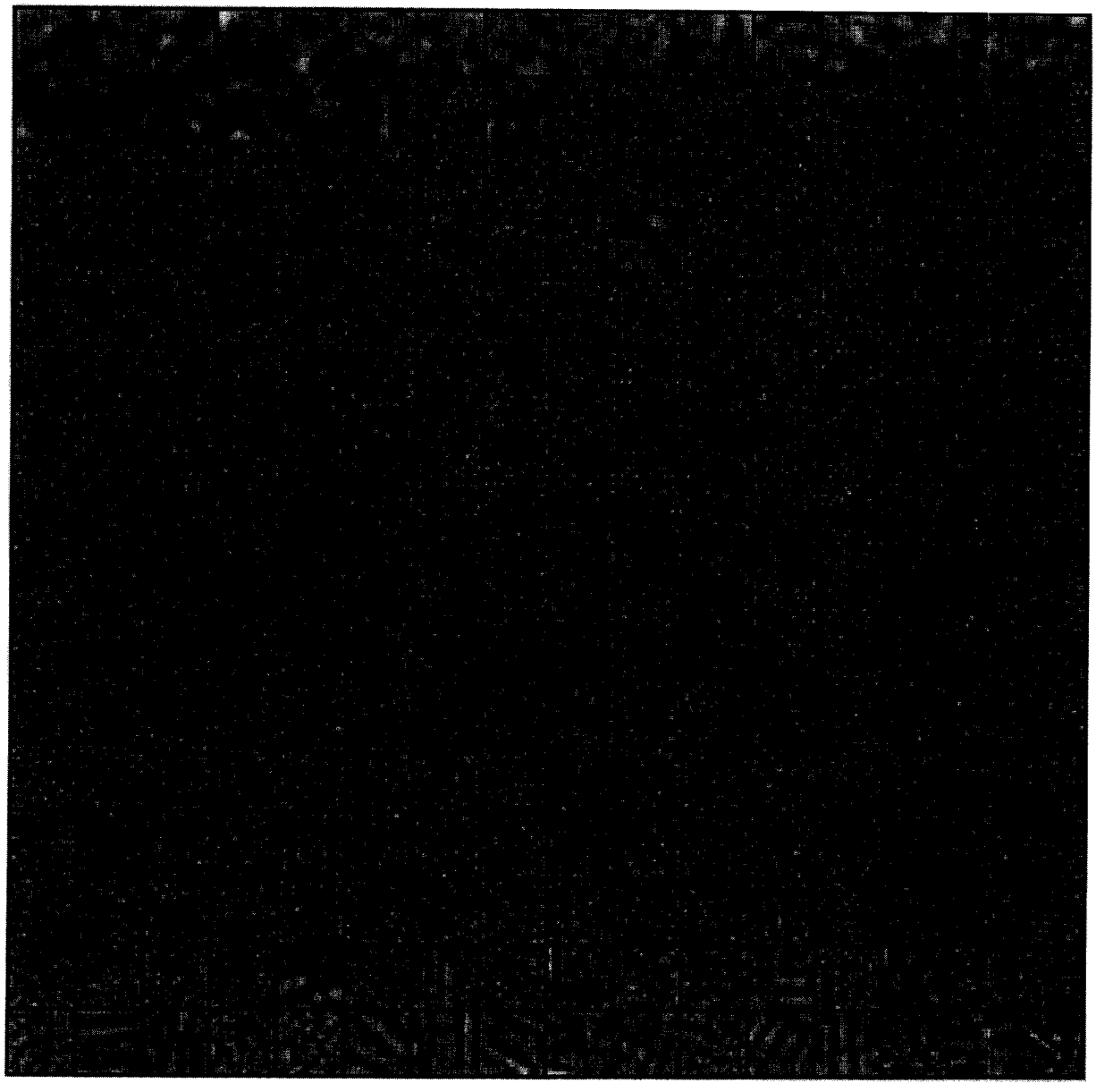


FIGURE 9. The basis B_{final} of receptive fields resulting from applying a biconvergent SDP to a training sequence of natural images. Receptive fields are ordered in terms of their average contribution (across all images in the test sequence) to the f_{final}-distortion computed in the structure test.

Results

Overall performance. H_0 was rejected by B_{final} in the structure test at the 0.005 level of significance, for 85.4% of the patches in the testing sequence. The average p-value over all patches in the testing sequence was 0.019. These results are comparable with those obtained using the non-biconvergent SDP in "*Discovering structure in natural images with an SDP*".

Resulting receptive fields B_{final} and histogram measure f_{final}. The Legendre code for histogram distortion template f_{final} (Fig. 10) assigns $w_1 = -0.0107$, $w_2 = -0.5070$, $w_3 = 0.0161$, $w_4 = 0.1945$, $w_5 = 0.0021$, $w_6 = 0.3191$, $w_7 = 0.0088$, $w_8 = 0.3502$, $w_9 = 0.0068$, $w_{10} = 0.3966$, $w_{11} = -0.0029$, and $w_{12} = 0.5682$, where, for $i = 1, 2, \ldots, 12$, w_i gives the coefficient of λ_i in the synthesis of f_{final}. The evident even symmetry of f_{final} is reflected by the fact that the coefficients of the odd-symmetric Legendre polynomials, $\lambda_1, \lambda_3, \ldots, \lambda_{11}$, are all near 0, whereas the coefficients of the even-symmetric components are relatively large in absolute value. The striking oscillations of the histogram distortion template f_{final} are an artifactual consequence of the truncation in our coding scheme for histogram distortion templates. f_{final} contains a large amount of λ_{12}, which has 13 extrema placed similarly to those of f_{final}. The high contribution to f_{final} of λ_{12} reflects the usefulness in rejecting H_0 of endowing f_{final} with steep, positive tails, and λ_{12} is the highest-number λ available.

The receptive fields of B_{final}, shown in Fig. 9, are ordered in terms of their average effectiveness in contributing to the rejection of H_0. Specifically, on each

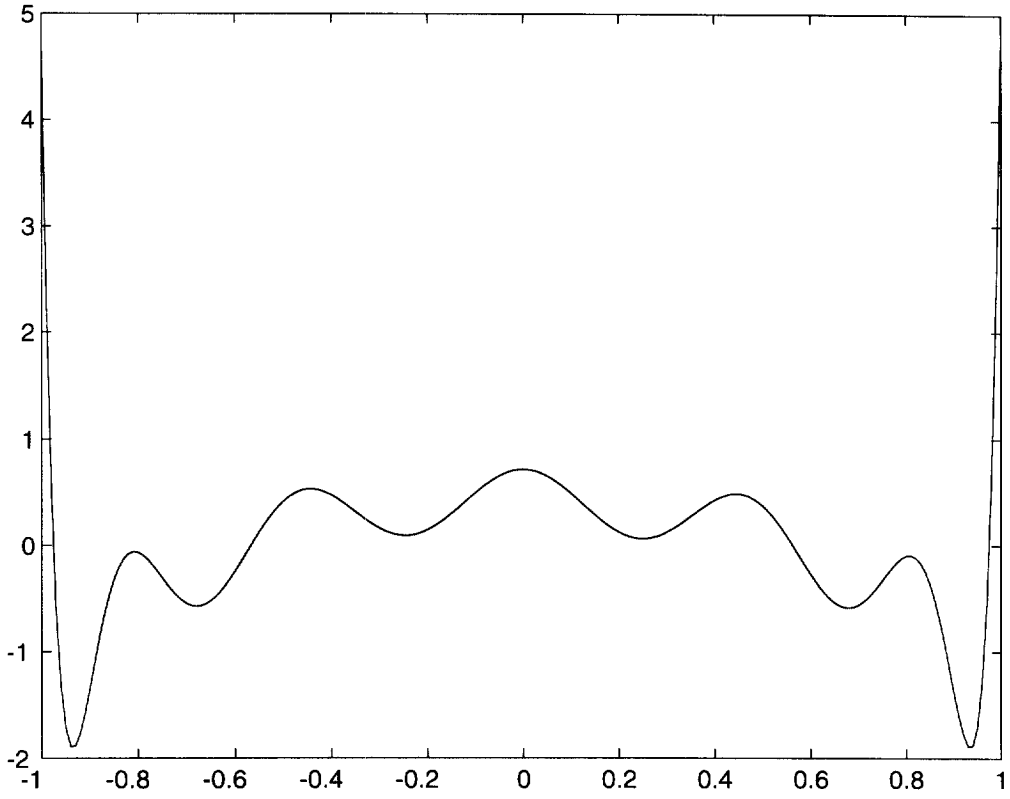


FIGURE 10. The histogram distortion template f_{final} resulting from applying a biconvergent SDP to a training sequence of natural images. Note the high values assigned near 1 and -1.

test trial i, for $u = 2\Phi \circ y - 1$, with $y = B_{\text{final}}g$, the quantity $f_{\text{final}}(u[j])$ was recorded for each receptive field j. $f_{\text{final}}(u[j])$ is the contribution of the j^{th} receptive field to the f_{final}-distortion of u. The greater $f_{\text{final}}(u[j])$ is, the more effective the j^{th} receptive field is in helping reject H_0 for test image x.

The receptive fields shown in Fig. 9 are ordered according to the average of $f_{\text{final}}(u[j])$ across all test images. As is evident, the most effective receptive fields are selective for low spatial frequencies; they do not, however, seem to be very well tuned for orientation. Due to the predominance of low spatial frequencies in natural images, these receptive fields tend to produce responses that deviate dramatically from 0. As a result, for such a receptive field j, the value $u[j]$ tends to occur very close either to 1 or else to -1. The histogram distortion template f_{final} has a pair of high-valued tails that provide sensitivity to these extreme values. Perhaps surprisingly, the least effective receptive fields (those occurring at the bottom of Fig. 9) seem to embody the highest degree of evident structure. The majority of apparently structureless receptive fields in the wide central region of Fig. 9 are more effective than the receptive fields at the bottom in rejecting H_0. The reason for this is that the receptive fields sandwiched in the center are actually high-frequency selective. Because natural images have little energy in the high frequency range, these receptive fields consistently give responses near 0. For such a receptive field j, the random variable $u[j]$ tends to take values very near 0. The histogram distortion template f_{final} has a local maximum at 0 which provides sensitivity to the responses of these receptive fields. By contrast, the nicely oriented receptive fields at the bottom of Fig. 9 are tuned to a band of spatial frequencies whose contribution to natural images is in the range of what might be expected under H_0. The smallest average contribution to rejection of H_0 (i.e., the smallest average value of $f_{\text{final}}(u[j])$) across all test images, was 0.023, given by the receptive field shown in the bottom-right corner of Fig. 9. The largest average contribution to rejection of H_0 was 2.20, given by the receptive field in the top-left corner of Fig. 9.

DISCUSSION

The simulation described in the previous section demonstrates that a biconvergent SDP can successfully derive, in tandem, a histogram distortion template along with an orthonormal basis of receptive fields that together constitute an effective tool for rejecting H_0 for natural images. Note that the histogram distortion template f_{final} assigns

1. Large positive values in the neighborhoods of -1 and 1, in its extreme tails; and
2. Positive values in the neighborhood of 0.

These two observations imply that response histograms produced by applying B_{final} to natural images do, indeed, tend to deviate from normality in being overly kurtotic.

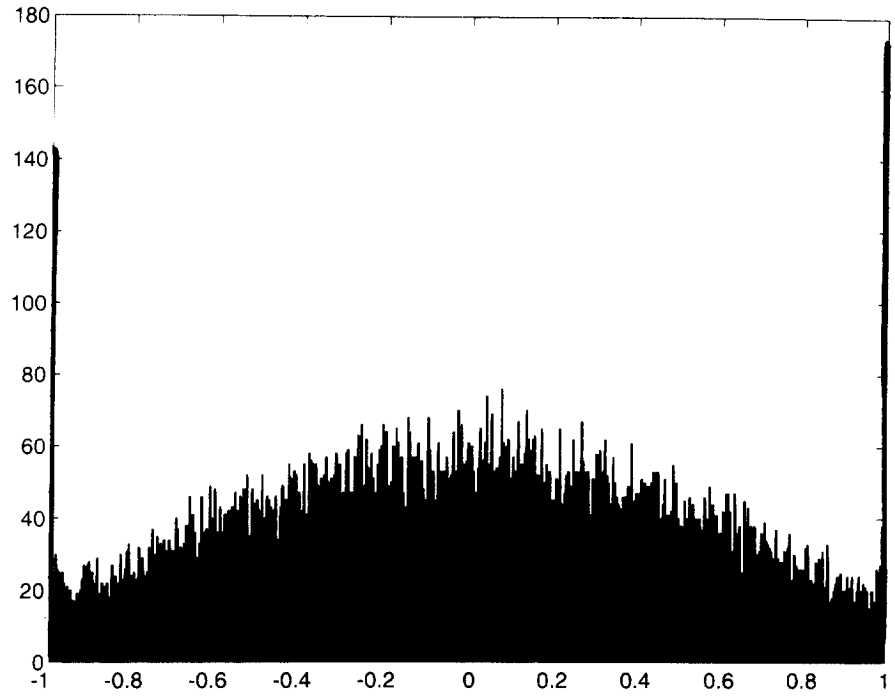


FIGURE 11. The histogram of receptive field responses obtained by applying the basis $B_{\text{non-bioconvergent}}$ (derived by the non-biconvergent SDP in the section entitled "*Discovering structure in natural images with an SDP*") to a sequence of 80, 16×16 patches of natural image. Responses r have been transformed into values $v = 2\Phi(r) - 1$. Under H_0, the transformed values v should be uniformly distributed on $(-1, 1)$. Note the very large spikes near -1 and 1, indicating that transformed responses v tend to take extreme values with high probability. Note also the similarity between this histogram and the histogram distortion template (Fig. 10) discovered by the biconvergent SDP.

The SDP implemented in '*Discovering structure in natural images with an SDP*' assumed a priori that response histograms of natural images will have high kurtosis. The biconvergent SDP was able to discover it.

This point is underscored by Fig. 11. Let $B_{\text{non-biconvergent}}$ denote the basis derived by the SDP applied to natural images in "*Discovering structure in natural images with an SDP*". To generate Fig. 11, each image x in a sequence of 80 natural image patches was transformed into $g = G(x)$; then $B_{\text{non-biconvergent}}$ was applied to g, and the vector of receptive field responses $y = B_{\text{non-biconvergent}}g$ was pointwise transformed to produce $u = 2\Phi \circ y - 1$. Under H_0, each vector u should consist of jointly independent random variables, all uniformly distributed on $(-1, 1)$. Figure 11 shows the histogram of all $20480 = 80 \times 256$ receptive field response values (coordinate values of the 80 vectors u). The primary deviation of this histogram from uniformity occurs in the tails, where we find extremely large spikes indicating high frequencies for values very close to -1 and 1. In particular, note the similarity between this histogram and the histogram distortion template (Fig. 10) discovered by the biconvergent SDP in the section "*A simulation using a biconvergent SDP*".

The reader will have noted that the receptive fields obtained by the biconvergent SDP (Fig. 9) do not seem as sharply refined as the receptive fields obtained with the non-biconvergent SDP. A comparison of Figs 10 and 11 suggests a possible explanation for this result. The Legendre coding scheme used by the biconvergent SDP severely limits the sensitivity to extreme values that it is possible for a histogram distortion template to achieve. It seems likely that these representational limitations may preclude any very sharp refinement of the corresponding basis of receptive fields.

Toward a more sensitive structure test

The structure test used by the biconvergent SDP could be made much more sensitive than it currently is. To see how, reflect that for some receptive fields j of B_{final}, the values $u[j]$, accumulated over all test images x, concentrate near 0—this is the case for the high-frequency receptive fields occurring in the middle of Fig. 9. For other receptive fields j, the values $u[j]$ concentrate near 1 and -1—this is the case for the low-frequency selective receptive fields occurring at the top of Fig. 9. Still other receptive fields j yield values $u[j]$ that accumulate according to less easily defined patterns—this is the case for the receptive fields at the bottom of Fig. 9. As these observations suggest, for any given receptive field j, there exists a histogram distortion template f_j that is uniquely sensitive to the histogram of response values $u[j]$ yielded by receptive field j. In effect, the current SDP implementation imposes the constraint that all receptive fields share the same histogram distortion template. Rather than computing $\Delta_f(u)$, however, one might define F to be a vector $(f_1, f_2, \ldots, f_N)$ of

histogram distortion templates, one for each receptive field, and set:

$$\Xi_F(u) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} f_j(u[j]), \tag{33}$$

where u comprises the pointwise-transformed receptive field outputs of the gaussian replacement of a given input image from the target population. Equation (33) generalizes equation (17) used to define the statistic Δ_f. Like Δ_f, Ξ_F has a standard normal distribution under H_0. It is clear, however, that for many image populations, Ξ_F can achieve much more power in rejecting H_0 than Δ_f. Thus, an important avenue of future research involves developing biconvergent SDPs that associate individual histogram distortion templates with different receptive fields.

REFERENCES

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

Barrow, H. G. (1987). Learning receptive fields. *IEEE First International Conference on Neural Networks, 4*, 115–121.

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

Diaconis, P. & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of Statistics, 12, (3)*, 793–815.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4*, 2379–2394.

Foldiak, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics, 64*, 165–170.

Fyfe, C. & Baddeley, R. (1995). Finding compact and sparse-distributed representations of visual images. *Network, 6*, 333–344.

Gonzalez, R. C. & Wintz, P. (1987). *Digital image processing*, 2nd edn. Reading, MA: Addison-Wesley.

Hancock, P. J. B., Baddeley, R. J. & Smith, L. S. (1992). The principal components of natural images. *Network, 3*, 61–72.

Harpur, G. F. & Prager, R. W. (1996). Development of low entropy coding in a recurrent network, *Network, 7*.

Hays, W. L. (1988). *Statistics*, New York: Holt, Reinhardt & Winston.

Huber, P. J. (1985). Projection Pursuit. *Annals of Statistics, 13*, 435–475.

Intrator, N. (1992). Feature extraction using an unsupervised neural network. *Neural Computation, 4*, 98–107.

Intrator, N. & Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions. *Neural Networks, 5*, 3–17.

Law, C. C. & Cooper, L. N. (1994). Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. *Proceeedings of the National Academy of Sciences USA, 91*, 7797–7801.

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

Liu, Y. & Shouval, H. (1994). Localized principal components of natural images—an analytic solution. *Network—Computation In Neural Systems, 5, (2)*, 317–324.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems, 1, (1)*, 61–68.

Olshausen, B. A. & Field, D. J. (1996). Natural images and efficient coding, *Network, 7*, 333–339.

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

PhotoDisc, Inc. (1995). *The starter kit*. 2013 Fourth Avenue, Seattle, Washington.

Ruderman, D. L. & Bialek, W. (1992). Seeing beyond the Nyquist limit. *Neural Computation, 4, (5)*, 682–690.

Ruderman, D. L. & Bialek, W. (1994). Statistics of natural images—scaling in the woods. *Physics Review Letters, 73, (6)*, 814–817.

Sanger, T. D. (1989,) An optimality principle for unsupervised learning. In Touretzky, D. S. (Ed.), *Advances in neural information processing systems* (pp. 11–19). San Mateo, CA: Morgan Kaufmann.

Schmidhuber, J., Eldracher, M. & Foltin, B. (1996). Semilinear predictability minimization produces well-known feature detectors, *Neural Computation, 8, (4)*, 773–786.

Shouval, H. & Liu, Y. (1996). Principal component neurons in a realistic visual environment. *Network—Computation In Neural Systems, 7 (3)*, 501–515.

Acknowledgement—Supported by AFOSR Life Sciences Visual Information Processing Program grant F49620-94-1-0345 and NSF Cognitive, Psychological and Language Sciences Program grant 9203291.